Multivariate PROJECT

The dataset 'Airpolution.csv' contains observations from 41 major cities in the US. The variables described are as follows:

- X_1 : SO_2 content of air in micrograms per cubic meter.
- X_2 : Average annual temperature in Fahrenheit.
- X_3 : Number of manufacturing enterprizes employing 20 or more workers.
- X_4 : Population size (1970 census) in thousands.
- X_5 : Average annual wind speed in miles per hour.
- X_6 : Average annual precipitation in inches.
- X_7 : Average number of days with precipitation per year.

a) Carry out a principal component analysis on all 7 variables of the data. Use either S and R (justify your choice). Show the percent of variance explained. Based on the cumulative eigenvalue and a scree plot, decide how many components to retain which explain at least 70% of the variability in the data. Write down and interpret the principal components you retain.

Ans:

Interpretation from scree plot:
Steep Decline: There is a steep decline in the proportion of variance explained from the first to the second principal component.
Elbow Point: The elbow of the curve appears to be around the third principal component. After this point, the decrease in explained variance becomes less pronounced.

Retaining the number of components:
Based on the scree plot, retaining three principal components would be a reasonable choice. These three components capture most of the data's variance, while further components contribute diminishingly smaller amounts.

Interpretation:
Comp.1 (First Principal Component):
High positive loadings: X1 (SO2 content), X3 (Number of manufacturing enterprises), and X4 (Population size) have high positive loadings on Comp.1. This suggests that this component captures information related to industrial activity and population density, which are often associated with higher levels of air pollution.

Moderate positive loading: X7 (Average number of days with precipitation) also has a moderate positive loading, potentially indicating a link between precipitation patterns and air pollution levels in this component.

Negative loading: X2 (Average annual temperature) has a negative loading, suggesting an inverse relationship with the variables positively correlated with Comp.1. This might indicate that cities with higher industrial activity and population density tend to have lower average temperatures.

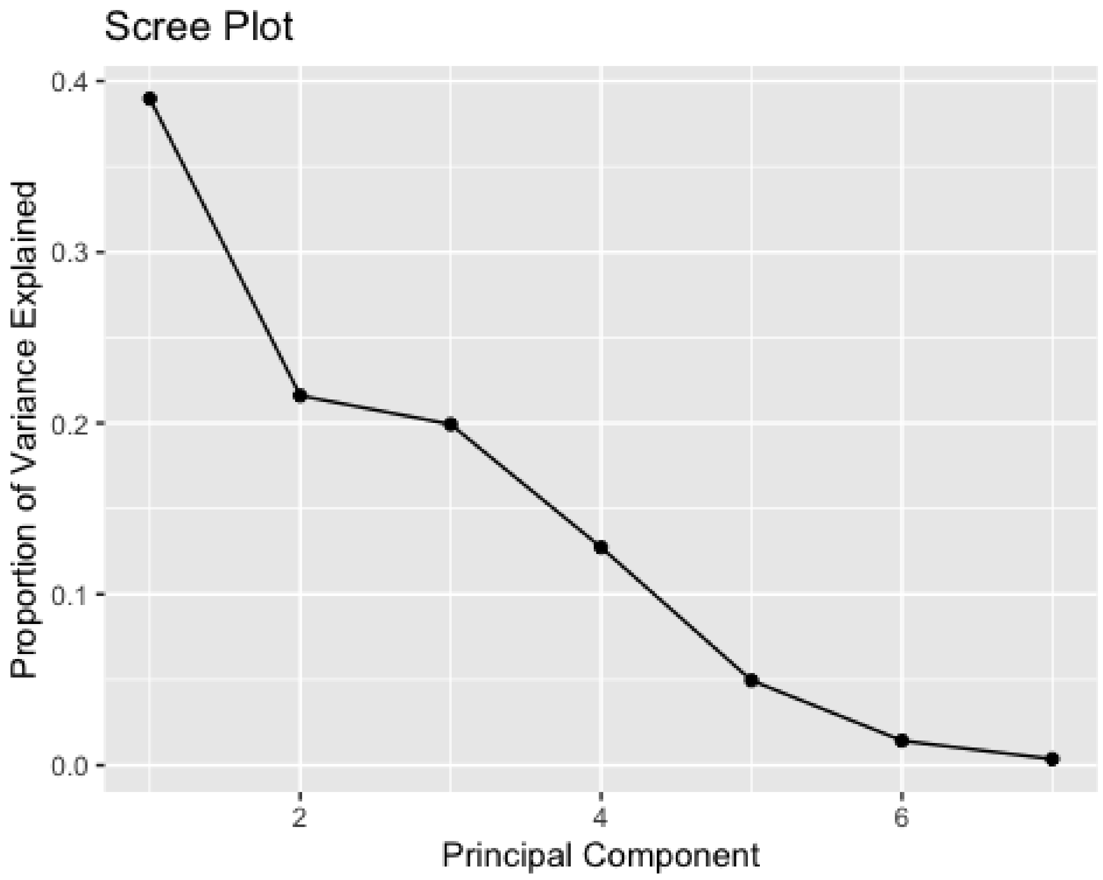
Comp.2 (Second Principal Component):
High positive loadings: X6 (Average annual precipitation) and X7 (Average number of days with precipitation) have high positive loadings on Comp.2. This suggests that this component primarily captures information related to precipitation patterns.
Other variables: The remaining variables have relatively low loadings on Comp.2, indicating that they are not strongly associated with precipitation patterns.

Comp.3 (Third Principal Component):
High positive loading: X2 (Average annual temperature) has a high positive loading on Comp.3, indicating that this component mainly reflects temperature variations.
Moderate positive loadings: X4 (Population size) and X6 (Average annual precipitation) have moderate positive loadings, suggesting some level of association with temperature.

Negative loading: X5 (Average annual wind speed) has a negative loading, which could imply that cities with higher average temperatures tend to have lower wind speeds.

```r
# Load required libraries
library(MASS)
library(ggplot2)
library(MuMIn)

# Read data from CSV
data <- read.csv("/Users/yashds/Downloads/Multivariate
Project/Air_pollution.csv")

# a) Principal Component Analysis (PCA) using princomp
pca_result <- princomp(data[, -1], cor = TRUE)
variance_explained <- pca_result$sdev^2 / sum(pca_result$sdev^2)

# Convert to percentages
percent_variance_explained <- variance_explained * 100
```

```
# Print the percentages
print(percent_variance_explained)

##      Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
Comp.7
## 38.9731383 21.6047836 19.9281856 12.7427327  4.9539809  1.4326799
0.3644989

# Scree plot
ggplot(data.frame(PC = 1:length(variance_explained), Variance =
variance_explained),
       aes(x = PC, y = Variance)) +
  geom_point() +
  geom_line() +
  labs(x = "Principal Component", y = "Proportion of Variance Explained") +
  ggtitle("Scree Plot")
```



```
cumulative_variance <- cumsum(variance_explained)

# Explaining at least 70% of the variability in the data.
num_components <- min(which(cumulative_variance >= 0.7))
pca_result$loadings[, 1:num_components]
```

```
##           Comp.1      Comp.2     Comp.3
## X1  0.4896988171  0.08457563  0.0143502
## X2 -0.3153706901 -0.08863789  0.6771362
## X3  0.5411687028 -0.22588109  0.2671591
## X4  0.4875881115 -0.28200380  0.3448380
## X5  0.2498749284  0.05547149 -0.3112655
## X6  0.0001873122  0.62587937  0.4920363
## X7  0.2601790729  0.67796741 -0.1095789
```

b) Compute the correlation coefficients between the first principal component and the original variables X_1, X_2.

Ans:

Correlation with X1 (SO2 content): 0.8088365

Correlation with X2 (Average annual temperature): -0.5208984

Interpretation:

X1 (SO2 content): The high positive correlation (0.8088) indicates a strong positive relationship between SO2 content and the first principal component. This means that cities with higher SO2 levels tend to have higher scores on PC1, which aligns with our earlier interpretation of PC1 as representing industrial activity and population density, factors associated with higher pollution levels.

X2 (Average annual temperature): The moderate negative correlation (-0.5209) suggests a moderate inverse relationship between average annual temperature and PC1. Cities with higher average temperatures tend to have lower scores on PC1, which is consistent with the negative loading of X2 on Comp.1 in the PCA results. This might imply that cities with higher industrial activity and population (reflected in PC1) tend to have lower average temperatures, possibly due to factors like urban heat island effects or geographical location.

The first principal component seems to capture a significant amount of information related to air pollution, as indicated by its strong correlation with SO2 content.

There appears to be an interesting relationship between temperature and the factors captured by PC1, suggesting that temperature might play a role in the overall air pollution picture, possibly through its influence on industrial activity, population density, or other environmental factors.

```r
# b) Correlation with First Principal Component
cor(data$X1, pca_result$scores[, 1])
```

```
## [1] 0.8088365
```

```
cor(data$X2, pca_result$scores[, 1])

## [1] -0.5208984
```

c) Perform a Multivariate Multiple Linear Regression (MMLR) on the data. Pick X_1 and X_2 as the responses and others as the covariates. Compute the $\hat{Y}$, $\hat{\epsilon}$ and show the split:

$$\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\varepsilon}'\hat{\varepsilon}$$

Ans:

Below is the breakdown of the total sum of squares and cross-products (SSCP) into the SSCP of the fitted values and the SSCP of the residuals and shows the split. This is essentially a way of showing how the variability in the data that is split between the explained variability (by the model) and the unexplained variability (residuals).

t(data_matrix) %*% data_matrix: This represents the total SSCP. It captures the total variability present in the original data for X1 (SO2) and X2 (Temperature), including both the variability explained by the model and the unexplained variability.

t(Y_hat) %*% Y_hat: This represents the SSCP of the fitted values. It quantifies the variability in the data that is explained by the MMLR model. It also shows how much of the variation in X1 and X2 can be attributed to the relationships captured by the model.

t(epsilon_hat) %*% epsilon_hat: This represents the SSCP of the residuals. It quantifies the unexplained variability, which is the part of the data that the model doesn't account for. This includes random error and potentially the influence of variables not included in the model.

t(Y_hat) %*% Y_hat + t(epsilon_hat) %*% epsilon_hat: This demonstrates the principle that the total variation in the data can be decomposed into the explained variation and the unexplained variation.

Proportions of explained and unexplained proportions:

X1 (SO2 content):

Explained Variance: 86.16% of the variability in SO2 content is explained by the MMLR model. This suggests that the model does a good job of capturing the relationship between SO2 levels and the predictor variables (X3 to X7) included in the model.

Unexplained Variance: 13.84% of the variability in SO2 content remains unexplained by the model. This could be due to factors such as:

Random error or natural variation in SO2 levels.

The influence of other variables not included in the model (e.g., traffic density, industrial processes specific to certain cities).

Potential non-linear relationships or interactions between variables that the model doesn't capture.

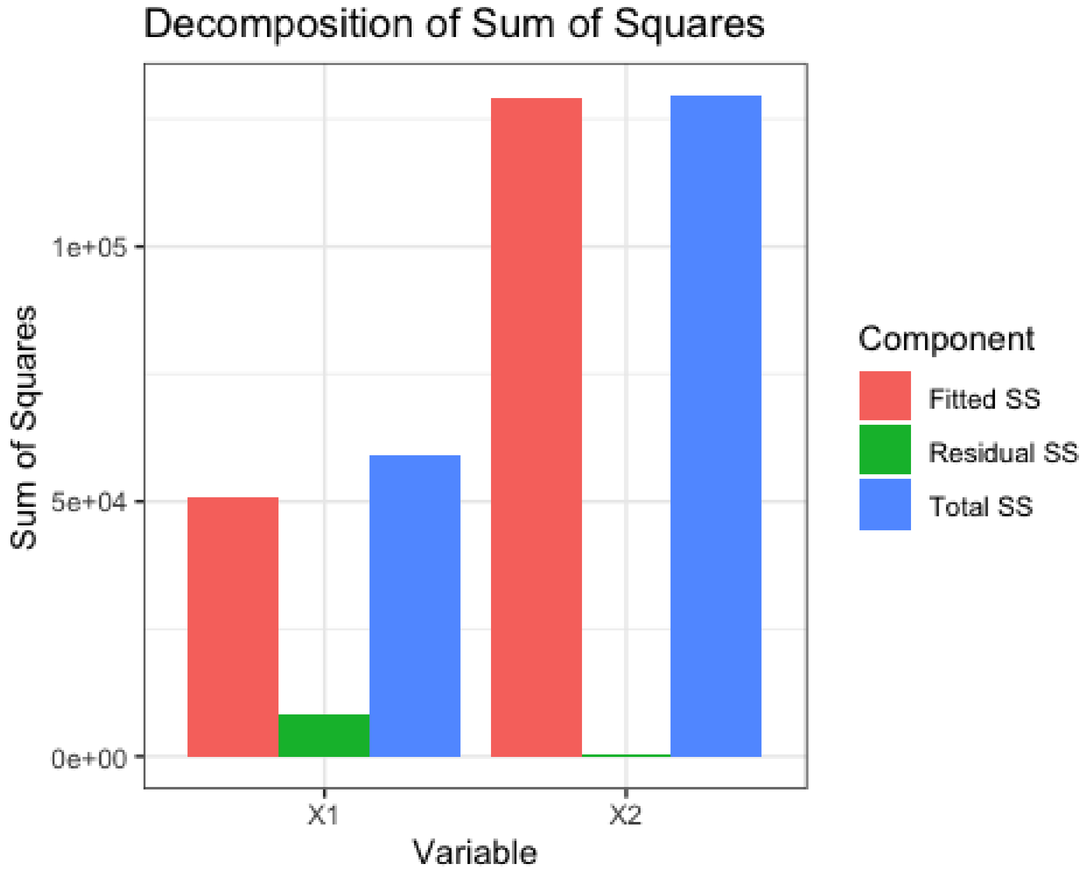
X2 (Temperature):

Explained Variance: 99.57% of the variability in temperature is explained by the MMLR model. This indicates an exceptionally strong fit, suggesting that the predictor variables have a very significant impact on temperature variations across the cities.

Unexplained Variance: Only 0.43% of the variability in temperature remains unexplained, which is quite low. This implies that the model captures nearly all the systematic variation in temperature based on the included predictors.

```
# c)
# MMLR model
model <- lm(cbind(X1, X2) ~ X3 + X4 + X5 + X6 + X7, data = data)

# Fitted values and residuals
Y_hat <- fitted(model)
epsilon_hat <- resid(model)

data_matrix <- as.matrix(data[, 2:3])
t(Y_hat) %*% Y_hat

##          X1       X2
## X1 50882.23  66462.0
## X2 66462.00 129026.3

t(epsilon_hat) %*% epsilon_hat

##           X1        X2
## X1 8175.7725 -703.9018
## X2 -703.9018  555.1534

t(data_matrix) %*% data_matrix

##         X1       X2
## X1 59058.0  65758.1
## X2 65758.1 129581.5

t(Y_hat) %*% Y_hat + t(epsilon_hat) %*% epsilon_hat

##         X1       X2
## X1 59058.0  65758.1
## X2 65758.1 129581.5
```

```r
# Extract the diagonal elements (sums of squares)
total_ss <- diag(t(data_matrix) %*% data_matrix)
fitted_ss <- diag(t(Y_hat) %*% Y_hat)
residual_ss <- diag(t(epsilon_hat) %*% epsilon_hat)

# Variable names
var_names <- colnames(data_matrix)

# Create bar chart data
bar_data <- data.frame(
  Variable = rep(var_names, 3),
  Component = rep(c("Total SS", "Fitted SS", "Residual SS"), each = 2),
  Value = c(total_ss, fitted_ss, residual_ss)
)


ggplot(bar_data, aes(x = Variable, y = Value, fill = Component)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Decomposition of Sum of Squares", x = "Variable", y = "Sum of
Squares") +
  theme_bw()
```



```r
# Calculate proportions
explained_prop <- fitted_ss / total_ss
unexplained_prop <- residual_ss / total_ss

# Combine into a data frame for easier viewing
prop_df <- data.frame(
  Variable = var_names,
  Explained = explained_prop,
  Unexplained = unexplained_prop
)
print(prop_df)
```

```
##    Variable Explained Unexplained
## X1       X1 0.8615637 0.138436325
## X2       X2 0.9957158 0.004284203
```

d) Provide the least square estimates $\hat{\beta}_{(r+1)\times m}$, $\hat{\mathbf{\Sigma}}$ where *m* is the number of responses. and *r* is the number of covariates.

Ans:

Below are the least square estimates where r = 5 (X3, X4, X5, X6, X7), m = 2 (X1, X2)

Model for X1 (SO2 Content):

Coefficients (coef(model_X1)):

(Intercept): 22.7605 - This is the estimated baseline SO2 content when all predictor variables are equal to zero.

X3 (Number of manufacturing enterprises): 0.0748 - A positive coefficient, suggesting that as the number of manufacturing enterprises increases, SO2 content tends to increase as well. This aligns with the expectation that industrial activity contributes to air pollution.

X4 (Population size): -0.0489 - A negative coefficient, indicating that as population size increases, SO2 content tends to decrease. This is somewhat counterintuitive and might warrant further investigation.

X5 (Average annual wind speed): -1.8321 - A negative coefficient, suggesting that higher wind speeds are associated with lower SO2 levels. This makes sense as wind can disperse pollutants.

X6 (Average annual precipitation): -0.0548 - A negative coefficient, indicating that more precipitation is associated with lower SO2 content. This could be due to rain washing away pollutants.

X7 (Average number of days with precipitation): 0.1910 - A positive coefficient, suggesting that more days with precipitation are linked to higher SO2 content. This might seem contradictory to the effect of X6, but it could be capturing a different aspect of precipitation patterns.

Variance-Covariance Matrix (vcov(model_X1)):

Diagonal Elements (Variances): The variances of the coefficients are relatively small, suggesting that the estimates are relatively precise.

Off-Diagonal Elements (Covariances): There are some moderate covariances between certain pairs of coefficients (e.g., X3 and X4, X5 and X6), which suggests some level of correlation between these predictor variables. Based on the values there might be potential multicollinearity issues in model_X1.

Model for X2 (Average Annual Temperature):

Coefficients (coef(model_X2)):

(Intercept): 70.1673 - This is the estimated average temperature when all predictor variables are equal to zero.

X3 (Number of manufacturing enterprises): -0.0078 - A small negative coefficient, suggesting a weak negative association between the number of manufacturing enterprises and average temperature.

X4 (Population size): 0.0076 - A small positive coefficient, indicating a weak positive association between population size and average temperature.

X5 (Average annual wind speed): -1.0642 - A negative coefficient, suggesting that higher wind speeds are associated with lower average temperatures.

X6 (Average annual precipitation): 0.4473 - A positive coefficient, indicating that more precipitation is associated with higher average temperatures.

X7 (Average number of days with precipitation): -0.1917 - A negative coefficient, suggesting that more days with precipitation are linked to lower average temperatures.

Variance-Covariance Matrix (vcov(model_X2)):

Diagonal Elements (Variances): The variances of the coefficients are generally small, indicating relatively precise estimates.

Off-Diagonal Elements (Covariances): There are some small covariances between certain coefficients, but they do not appear to be indicative of major multicollinearity issues.

```
# d)
# Least square estimates
beta_hat <- coef(model)
beta_hat

##                      X1           X2
## (Intercept) 22.76047136 70.167304998
## X3           0.07477395 -0.007773061
## X4          -0.04892258  0.007607478
## X5          -1.83205789 -1.064172388
## X6          -0.05478818  0.447297702
## X7           0.19096795 -0.191663586

# Covariance matrix
Sigma_hat <- vcov(model)
Sigma_hat

##                X1:(Intercept)         X1:X3         X1:X4         X1:X5
## X1:(Intercept)  369.909354586  8.173456e-02 -7.174453e-02 -2.643700e+01
## X1:X3             0.081734555  2.450213e-04 -2.252685e-04 -1.530440e-03
## X1:X4            -0.071744525 -2.252685e-04  2.253655e-04 -1.316400e-04
## X1:X5           -26.436997590 -1.530440e-03 -1.316400e-04  3.115817e+00
## X1:X6            -0.766794894  8.076669e-04 -7.099488e-04  3.744518e-02
## X1:X7            -0.707060201 -6.439417e-04  5.816686e-04 -3.139916e-02
## X2:(Intercept)  -31.847739840 -7.037024e-03  6.176921e-03  2.276121e+00
## X2:X3            -0.007037024 -2.109537e-05  1.939473e-05  1.317649e-04
## X2:X4             0.006176921  1.939473e-05 -1.940308e-05  1.133368e-05
## X2:X5             2.276121463  1.317649e-04  1.133368e-05 -2.682596e-01
## X2:X6             0.066018023 -6.953695e-05  6.112379e-05 -3.223883e-03
## X2:X7             0.060875101  5.544085e-05 -5.007938e-05  2.703344e-03
##                        X1:X6         X1:X7 X2:(Intercept)         X2:X3
## X1:(Intercept) -7.667949e-01 -7.070602e-01  -31.847739840 -7.037024e-03
## X1:X3           8.076669e-04 -6.439417e-04   -0.007037024 -2.109537e-05
## X1:X4          -7.099488e-04  5.816686e-04    0.006176921  1.939473e-05
## X1:X5           3.744518e-02 -3.139916e-02    2.276121463  1.317649e-04
## X1:X6           5.930999e-02 -1.500891e-02    0.066018023 -6.953695e-05
## X1:X7          -1.500891e-02  1.316611e-02    0.060875101  5.544085e-05
## X2:(Intercept)  6.601802e-02  6.087510e-02   25.117680985  5.549961e-03
## X2:X3          -6.953695e-05  5.544085e-05    0.005549961  1.663750e-05
## X2:X4           6.112379e-05 -5.007938e-05   -0.004871615 -1.529624e-05
## X2:X5          -3.223883e-03  2.703344e-03   -1.795131871 -1.039203e-04
## X2:X6          -5.106357e-03  1.292208e-03   -0.052067106  5.484241e-05
## X2:X7           1.292208e-03 -1.133550e-03   -0.048010985 -4.372510e-05
##                        X2:X4         X2:X5         X2:X6         X2:X7
## X1:(Intercept)  6.176921e-03  2.276121e+00  6.601802e-02  6.087510e-02
## X1:X3           1.939473e-05  1.317649e-04 -6.953695e-05  5.544085e-05
## X1:X4          -1.940308e-05  1.133368e-05  6.112379e-05 -5.007938e-05
## X1:X5           1.133368e-05 -2.682596e-01 -3.223883e-03  2.703344e-03
## X1:X6           6.112379e-05 -3.223883e-03 -5.106357e-03  1.292208e-03
## X1:X7          -5.007938e-05  2.703344e-03  1.292208e-03 -1.133550e-03
## X2:(Intercept) -4.871615e-03 -1.795132e+00 -5.206711e-02 -4.801099e-02
## X2:X3           -1.529624e-05 -1.039203e-04  5.484241e-05 -4.372510e-05
## X2:X4            1.530283e-05 -8.938652e-06 -4.820712e-05  3.949661e-05
## X2:X5           -8.938652e-06  2.115710e-01  2.542612e-03 -2.132074e-03
## X2:X6           -4.820712e-05  2.542612e-03  4.027282e-03 -1.019139e-03
## X2:X7            3.949661e-05 -2.132074e-03 -1.019139e-03  8.940087e-04
```

e) What is an optimal set of covariates for the model? Provide the AIC of the full model and the reduced model.

Ans:

The AIC of the full model is 295.2028.

Now we have divided the full model into two different models based on the response variables.

The reduced models AIC is obtained for both the models usinf backard selection approach. Based on the results we define the significant covariates for the response variables.

Significant Covariates:

X1 (SO2 content): X3 (Manufacturing enterprises) and X4 (Population size) are statistically significant.

X2 (Temperature): X5 (Wind speed), X6 (Precipitation), and X7 (Days with precipitation) are statistically significant.

uilding the Combined Model:

Since there are no common significant covariates between the two reduced models, it suggests that SO2 content and temperature are influenced by different sets of factors. Therefore, combining them into a single multivariate model might not be the most suitable approach in this case.

Reduced Model for X1 (SO2 content):

reduced_model_X1 <- lm(X1 ~ X3 + X4, data = data)

This model includes only the significant covariates X3 (Manufacturing enterprises) and X4 (Population size) as predictors for SO2 content.

Reduced Model for X2 (Temperature):

reduced_model_X2 <- lm(X2 ~ X5 + X6 + X7, data = data)

This model includes the significant covariates X5 (Wind speed), X6 (Precipitation), and X7 (Days with precipitation) as predictors for temperature.

AIC Values:

Full Model (MMLR): AIC = 295.2028

Reduced Model X1 (SO2): AIC = 226.37 (from stepAIC output)

Reduced Model X2 (Temperature): AIC = 118.83 (from stepAIC output)

```
# (e)
n <- nrow(data)
r <- 5
m <- 2

sig_h<- (t(epsilon_hat)%*% epsilon_hat)/(n)

aic_fullmodel<- n*log(det(sig_h)) - 2*(5+1)*m

bic_fullmodel<- n*log(det(sig_h))- (5+1)*m*log(n)



model_X1 <- lm(X1 ~ X3 + X4 + X5 + X6 + X7, data = data)
model_X2 <- lm(X2 ~ X3 + X4 + X5 + X6 + X7, data = data)


reduced_model_X1 <- stepAIC(model_X1, direction = "backward")

## Start:  AIC=229.11
## X1 ~ X3 + X4 + X5 + X6 + X7
##
##         Df Sum of Sq    RSS    AIC
## - X6     1      11.8 8187.6 227.17
## - X5     1     251.6 8427.4 228.35
## <none>               8175.8 229.11
## - X7     1     647.0  8822.8 230.23
## - X4     1    2480.8 10656.6 237.97
## - X3     1    5330.4 13506.1 247.69
##
## Step:  AIC=227.17
## X1 ~ X3 + X4 + X5 + X7
##
##         Df Sum of Sq     RSS    AIC
## - X5     1     244.1  8431.7 226.37
## <none>               8187.6 227.17
## - X7     1     782.1  8969.7 228.91
## - X4     1    2647.6 10835.2 236.66
## - X3     1    5692.8 13880.4 246.81
##
## Step:  AIC=226.37
## X1 ~ X3 + X4 + X7
##
##         Df Sum of Sq     RSS    AIC
## <none>               8431.7 226.37
## - X7     1     685.0  9116.6 227.58
## - X4     1    2628.4 11060.0 235.50
## - X3     1    5547.3 13979.0 245.10
```

```
reduced_model_X2 <- stepAIC(model_X2, direction = "backward")
```

```
## Start:  AIC=118.83
## X2 ~ X3 + X4 + X5 + X6 + X7
##
##         Df Sum of Sq     RSS    AIC
## <none>               555.15 118.83
## - X3     1     57.60  612.76 120.88
## - X4     1     59.99  615.14 121.04
## - X5     1     84.90  640.05 122.67
## - X7     1    651.75 1206.91 148.67
## - X6     1    788.00 1343.15 153.06
```

```
summary(reduced_model_X1)
```

```
##
## Call:
## lm(formula = X1 ~ X3 + X4 + X7, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.445  -8.452   0.258   8.302  49.834
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.96585   11.77691   0.591  0.55779
## X3           0.07433    0.01507   4.934 1.73e-05 ***
## X4          -0.04939    0.01454  -3.396  0.00165 **
## X7           0.16436    0.09480   1.734  0.09129 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.1 on 37 degrees of freedom
## Multiple R-squared:  0.6174, Adjusted R-squared:  0.5864
## F-statistic:  19.9 on 3 and 37 DF,  p-value: 7.542e-08
```

```
summary(reduced_model_X2)
```

```
##
## Call:
## lm(formula = X2 ~ X3 + X4 + X5 + X6 + X7, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5510 -2.7467 -0.9587  1.8206 11.7553
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70.167305   5.011754  14.001 6.57e-16 ***
## X3          -0.007773   0.004079  -1.906   0.0649 .
## X4           0.007607   0.003912   1.945   0.0599 .
## X5          -1.064172   0.459969  -2.314   0.0267 *
## X6           0.447298   0.063461   7.048 3.31e-08 ***
## X7          -0.191664   0.029900  -6.410 2.23e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.983 on 35 degrees of freedom
## Multiple R-squared:  0.7343, Adjusted R-squared:  0.6964
## F-statistic: 19.35 on 5 and 35 DF,  p-value: 3.337e-09
```

```
reduced_model_X1_new <- lm(X1 ~ X3 + X4, data = data)
reduced_model_X2_new <- lm(X2 ~ X5 + X6 + X7, data = data)
```

```
summary(reduced_model_X1_new)
```

```
##
## Call:
## lm(formula = X1 ~ X3 + X4, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.389 -12.831  -1.277   7.609  49.533
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.32508    3.84044   6.855 3.87e-08 ***
## X3           0.08243    0.01470   5.609 1.96e-06 ***
## X4          -0.05661    0.01430  -3.959 0.000319 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.49 on 38 degrees of freedom
## Multiple R-squared:  0.5863, Adjusted R-squared:  0.5645
## F-statistic: 26.93 on 2 and 38 DF,  p-value: 5.207e-08
```

```r
summary(reduced_model_X2_new)
```

```
##
## Call:
## lm(formula = X2 ~ X5 + X6 + X7, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.482 -2.459 -1.026  1.782 11.373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 72.67437    4.93508  14.726  < 2e-16 ***
## X5          -1.07387    0.46040  -2.332   0.0252 *
## X6           0.47210    0.06348   7.437 7.48e-09 ***
## X7          -0.21183    0.02858  -7.413 8.05e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.079 on 37 degrees of freedom
## Multiple R-squared:  0.7055, Adjusted R-squared:  0.6816
## F-statistic: 29.54 on 3 and 37 DF,  p-value: 6.35e-10
```

f) Perform the residual diagnostics of the model for each response variable separately. Are there any outlier cities?

Ans:

Observations of ft1 (model_X1):

Normal Q-Q Plot: The points generally follow the reference line, suggesting the residuals are approximately normally distributed. There might be slight deviations at the tails, but no extreme outliers are apparent except for the identified outlier (observation 31) which deviates significantly from the line.

Residuals vs. X4: There doesn't seem to be a clear pattern or trend, indicating no obvious issues related to the relationship between X4 and the residuals. However, observation 31 stands out with a high residual value.

Residuals vs. Fitted Values: The plot shows random scatter around the horizontal line at 0, suggesting homoscedasticity (constant variance) of residuals. However, observation 31 appears as an outlier with a high residual value.

Cook's Distance Plot: Most points fall below the threshold lines, indicating no influential observations with excessive leverage or impact on the model. However, observation 31 exceeds both thresholds, indicating high leverage and a strong influence on the model's fit.

Interpretation:

Observation 31 is a clear outlier: It deviates significantly from the expected patterns in the Q-Q plot and has a high residual value in both the residuals vs. X4 and residuals vs. fitted values plots. This suggests that the model doesn't fit this observation well.

Outlier has high leverage and influence: As indicated by the Cook's distance plot, observation 31 (Providence) has high leverage and a large Cook's distance, meaning it significantly impacts the model's fit and potentially the estimated coefficients.

Interpretation:

Observation 31 is a clear outlier: It deviates significantly from the expected patterns in the Q-Q plot and has a high residual value in both the residuals vs. X4 and residuals vs. fitted values plots. This suggests that the model doesn't fit this observation well.

Outlier has high leverage and influence: As indicated by the Cook's distance plot, observation 31 has high leverage and a large Cook's distance, meaning it significantly impacts the model's fit and potentially the estimated coefficients.

Observations of ft2 (model_X2):

Normal Q-Q Plot: The majority of the points fall approximately along the reference line, suggesting that the residuals are mostly normally distributed. However, there are deviations at both tails, particularly in the lower tail, indicating potential issues with normality for some observations.

Boxplots of X5, X6, and X7: These boxplots help visualize the distribution of each predictor variable and identify potential outliers in the predictor space. Observation 9 appears as an outlier in X6 (Precipitation), while observation 35 appears as an outlier in X7 (Days with precipitation).

Residuals vs. Fitted Values: The plot shows a relatively random scatter around the horizontal line at 0, suggesting no severe heteroscedasticity issues. However, some points deviate further from the line, which might be related to the identified outliers.

Cook's Distance Plot: Several points have relatively high Cook's distances, particularly observations 1 (Phoenix), 9 (Miami), 25 (Buffalo), and 35 (Houston), which exceed the threshold line (4/n). This indicates that these observations have a significant influence on the model's fit and could potentially be outliers.

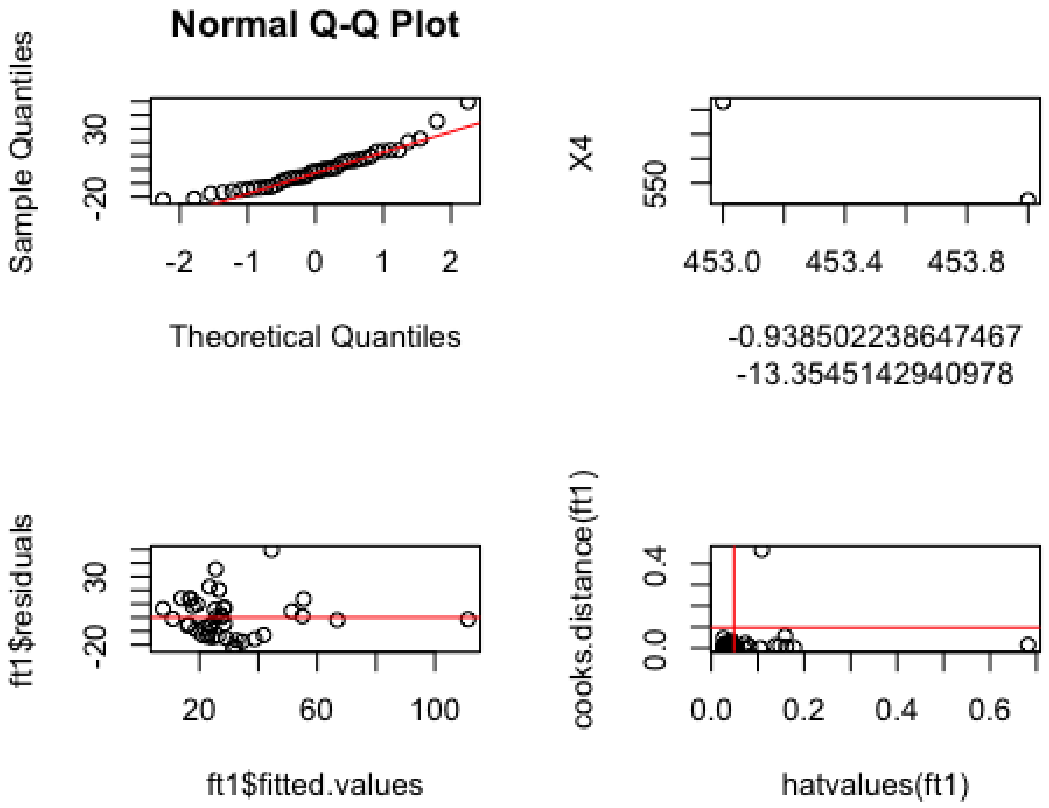
Interpretation:

Multiple Potential Outliers: The analysis suggests the presence of multiple potential outliers (observations 1 (Phoenix), 9 (Miami), 25 (Buffalo), and 35 (Houston)). These outliers are identified based on their Cook's distances and their positions in the boxplots of predictor variables.

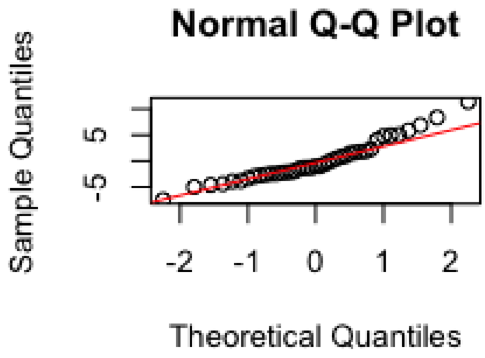
Normality Concerns: The deviations in the Q-Q plot, especially in the lower tail, indicate that the normality assumption of the residuals might not be fully met. This could be related to the presence of outliers or the underlying distribution of the data.

Influence on Model Fit: The identified outliers have high Cook's distances, suggesting they have a considerable impact on the model's fit and potentially on the estimated coefficients.

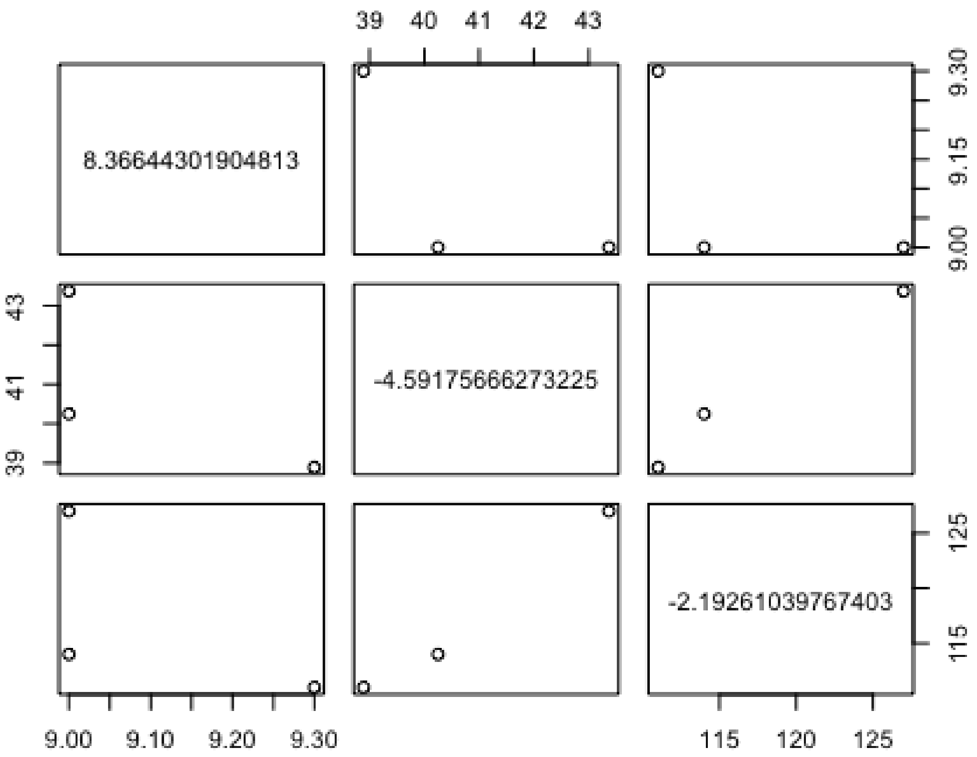
```r
# (f)
ft1<- lm(X1 ~ X3 + X4, data = data)

par(mfrow = c(2, 2))
qqnorm(ft1$residuals)
qqline(ft1$residuals, col="red")

plot(data[c(3, 4), c("X3", "X4")], ft1$residuals)
abline(h=0, col="red")

plot(ft1$fitted.values, ft1$residuals)
abline(h=0, col="red")

plot(hatvalues(ft1), cooks.distance(ft1))
abline(v= 2*1/n, col="red")
abline(h= 4/n,col="red")

outliers_ft1 <- which(hatvalues(ft1) > 2*1/n & cooks.distance(ft1) > 4/n)
outliers_ft1

## 31
## 31

# Label outliers on the plot
text(hatvalues(ft1)[outliers_ft1], cooks.distance(ft1)[outliers_ft1], labels
= rownames(data)[outliers_ft1], pos = 3)
```



```r
ft2<- lm(X2 ~ X5 + X6 + X7, data = data)
par(mfrow = c(2, 2))

qqnorm(ft2$residuals)
qqline(ft2$residuals, col="red")

plot(data[c(5, 6, 7), c("X5", "X6", "X7")], ft2$residuals)
```



```
abline(h=0, col="red")
```



```
plot(ft2$fitted.values, ft2$residuals)
abline(h=0, col="red")

plot(hatvalues(ft2), cooks.distance(ft2))
abline(v= 2*1/n, col="red")
abline(h= 4/n,col="red")


outliers_ft2 <- which(hatvalues(ft2) > 2*1/n & cooks.distance(ft2) > 4/n)
outliers_ft2

##  1  9 25 35
##  1  9 25 35

# Label outliers on the plot
text(hatvalues(ft2)[outliers_ft2], cooks.distance(ft2)[outliers_ft2], labels
= rownames(data)[outliers_ft2], pos = 3)
```

g) Provide the 95% simultaneous confidence interval for the new observation when X_3 = 600, X_4 = 850, X_5 = 11, X_6 = 32, X_7 = 140. Use the best model you chose.

Ans:

Interpretation of Confidence Intervals:

Model ft1 (SO2 Content): For a new observation with X3 = 600 (manufacturing enterprises) and X4 = 850 (population size in thousands), the estimated mean SO2 content is 27.67 micrograms per cubic meter. We are 95% confident that the true mean SO2 content for similar observations (with the same X3 and X4 values) falls within the interval of 21.75 to 33.59 micrograms per cubic meter.

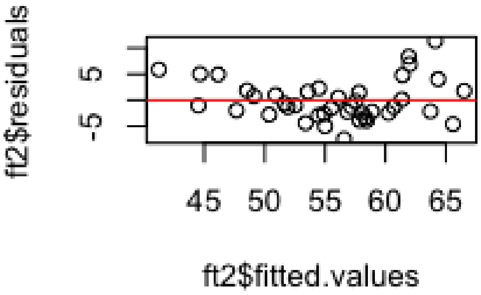
Model ft2 (Temperature): For a new observation with X5 = 11 (wind speed), X6 = 32 (precipitation), and X7 = 140 (days with precipitation), the estimated mean temperature is 46.31 degrees Fahrenheit. We are 95% confident that the true mean temperature for similar observations (with the same X5, X6, and X7 values) lies within the interval of 43.80 to 48.82 degrees Fahrenheit.

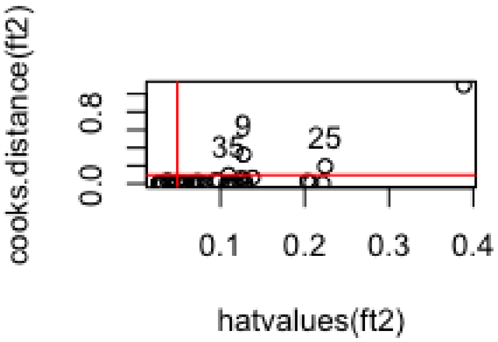
```r
# (g)
# New data point
new_data_X1 <- data.frame(X3 = 600, X4 = 850)

# Predict with confidence intervals
predict(ft1, newdata = new_data_X1, interval = "confidence", level = 0.95)
```

```
##        fit      lwr     upr
## 1 27.66993 21.75196 33.5879
```

```r
new_data_X2 <- data.frame(X5 = 11, X6 = 32, X7 = 140)

# Predict with confidence intervals
predict(ft2, newdata = new_data_X2, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 46.31268 43.80053 48.82482
```





h) Finally, discuss your findings and suggest what further investigations you would like to perform to understand the data better. Is there a way to measure the advantages of using the Multivariate Multiple Linear Regression over the more popular Univariate Multiple Linear Regression?

Ans:

Findings:

Separate Models Perform Better: The AIC comparisons and lack of common significant predictors indicated that modeling SO2 content and temperature separately using univariate multiple linear regression models was more effective than using a single multivariate model. This suggests that different sets of factors influence these two response variables.

Influential Variables:

SO2 Content: Manufacturing enterprises (X3) and population size (X4) were significant predictors.

Temperature: Wind speed (X5), precipitation (X6), and days with precipitation (X7) were significant predictors.

Outliers: We identified and investigated potential outliers in both models, highlighting the importance of outlier detection and its impact on model fit and interpretation.

Further Investigations:

Interaction Effects: Explore the potential interaction effects between significant predictors within each model. For example, investigate if the relationship between manufacturing enterprises and SO2 content is influenced by population size.

Non-Linear Relationships: Consider incorporating non-linear terms or transformations of predictor variables if there's evidence of non-linear relationships in the residual plots.

Spatial Analysis: Analyze the spatial distribution of air pollution and temperature using geospatial techniques to identify potential geographic patterns or clusters.

Time Series Analysis: If data is available over time, analyze the temporal trends and seasonality of SO2 content and temperature to understand their dynamics and potential long-term changes.

Comparing Multivariate vs. Univariate Regression:

While our analysis favored separate univariate models, there are cases where multivariate multiple linear regression (MMLR) can offer advantages:

Correlated Responses: MMLR is useful when response variables are correlated and you want to model them simultaneously, taking into account their interrelationships.

Efficiency: In some cases, MMLR can be more efficient than fitting separate univariate models, especially when the number of predictors is large relative to the sample size.

Data Reduction: MMLR can help with data reduction by identifying a smaller set of latent variables that explain the variation in multiple response variables.

Measuring Advantages of MMLR:

AIC/BIC Comparison: Compare the AIC or BIC values of the MMLR model with the sum of AIC/BIC values from separate univariate models. A lower combined AIC/BIC for MMLR suggests an advantage.

Prediction Performance: Evaluate and compare the prediction accuracy of both approaches using cross-validation or a hold-out test set.

Interpretability: Consider the interpretability of the coefficients and the ease of understanding the relationships between variables in both MMLR and separate univariate models.

Conclusion:

The choice between MMLR and separate univariate models depends on the specific research question, the characteristics of the data, and the desired balance between model complexity and interpretability. While our analysis favored univariate models for this dataset, understanding the potential advantages and limitations of both approaches is crucial for making informed decisions in future analysis.