STA521_Lab03_2024Fall

Justin Kao

2024-09-23

Obtain the data and manipulate it into the appropriate format.

Load Data

```
data("Ozone", package = "mlbench")
```

Data Wrangling

```
names(Ozone) <- c("Month", "Day_of_Month", "Day_of_Week", "Ozone",
                        "Pressure", "Wind", "Humidity", "Temp1", "Temp2",
                        "Inversion_Height", "Pressure_Gradient",
                        "Inversion_Temp", "Visibility")

Ozone <- na.omit(Ozone)


nrow(Ozone)
```

```
## [1] 203
```

```
Ozone$Month <- as.numeric(Ozone$Month)
Ozone <- Ozone[, !(names(Ozone) %in% c("Day_of_Month", "Day_of_Week"))]
str(Ozone)
```

```
## 'data.frame':    203 obs. of  11 variables:
##  $ Month            : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Ozone            : num  5 6 4 4 6 6 5 4 4 7 ...
##  $ Pressure         : num  5760 5720 5790 5790 5700 5720 5760 5780 5830 5870 ...
##  $ Wind             : num  3 4 6 3 3 3 6 6 3 2 ...
##  $ Humidity         : num  51 69 19 25 73 44 33 19 19 19 ...
##  $ Temp1            : num  54 35 45 55 41 51 51 54 58 61 ...
##  $ Temp2            : num  45.3 49.6 46.4 52.7 48 ...
##  $ Inversion_Height : num  1450 1568 2631 554 2083 ...
##  $ Pressure_Gradient: num  25 15 -33 -28 23 9 -44 -44 -53 -67 ...
##  $ Inversion_Temp   : num  57 53.8 54.1 64.8 52.5 ...
##  $ Visibility       : num  60 60 100 250 120 150 40 200 250 200 ...
```

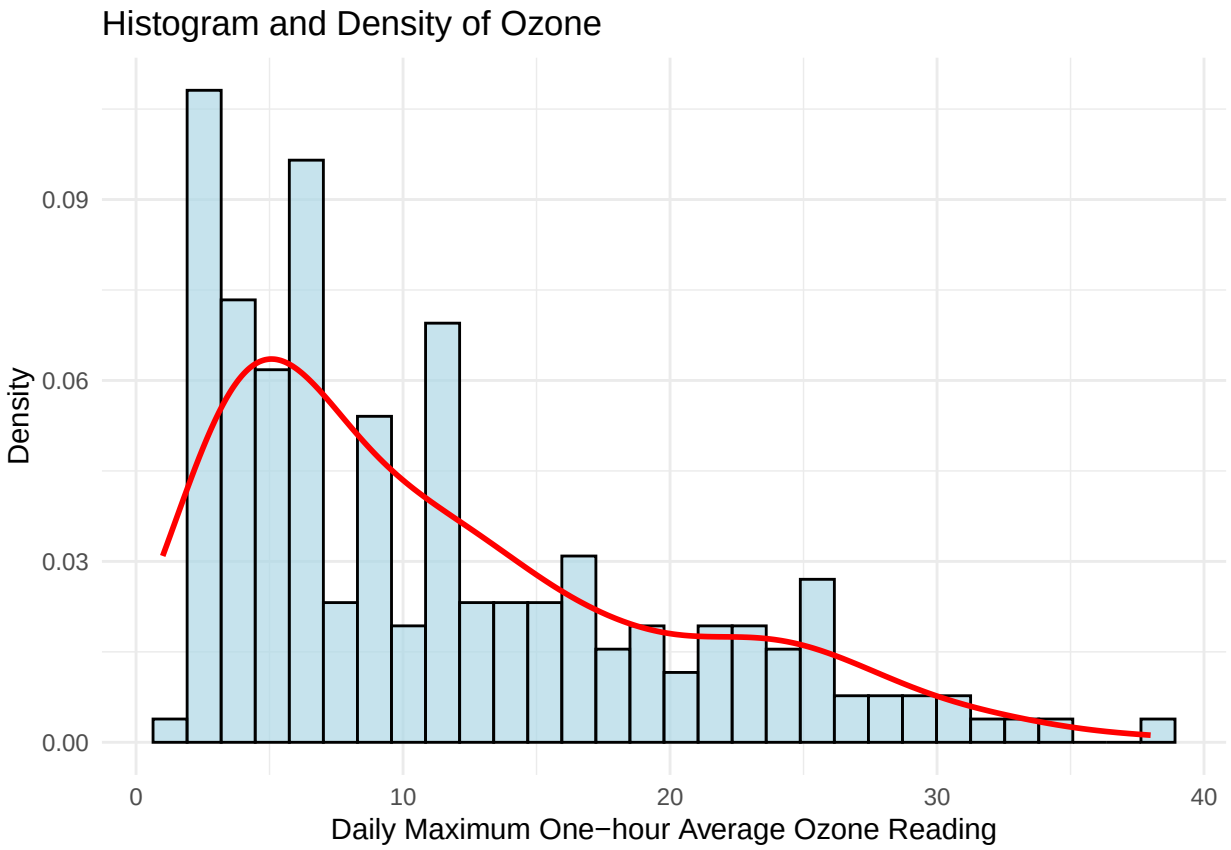
Basic Plots

```
library(ggplot2)
ggplot(Ozone, aes(x = Ozone)) +
  geom_histogram(aes(y = ..density..), bins = 30, color = "black", fill = "lightblue", alpha = 0.7) +
  geom_density(color = "red", size = 1) +
```

```
  labs(x = "Daily Maximum One-hour Average Ozone Reading", y = "Density", title = "Histogram and Density
  theme_minimal()
```

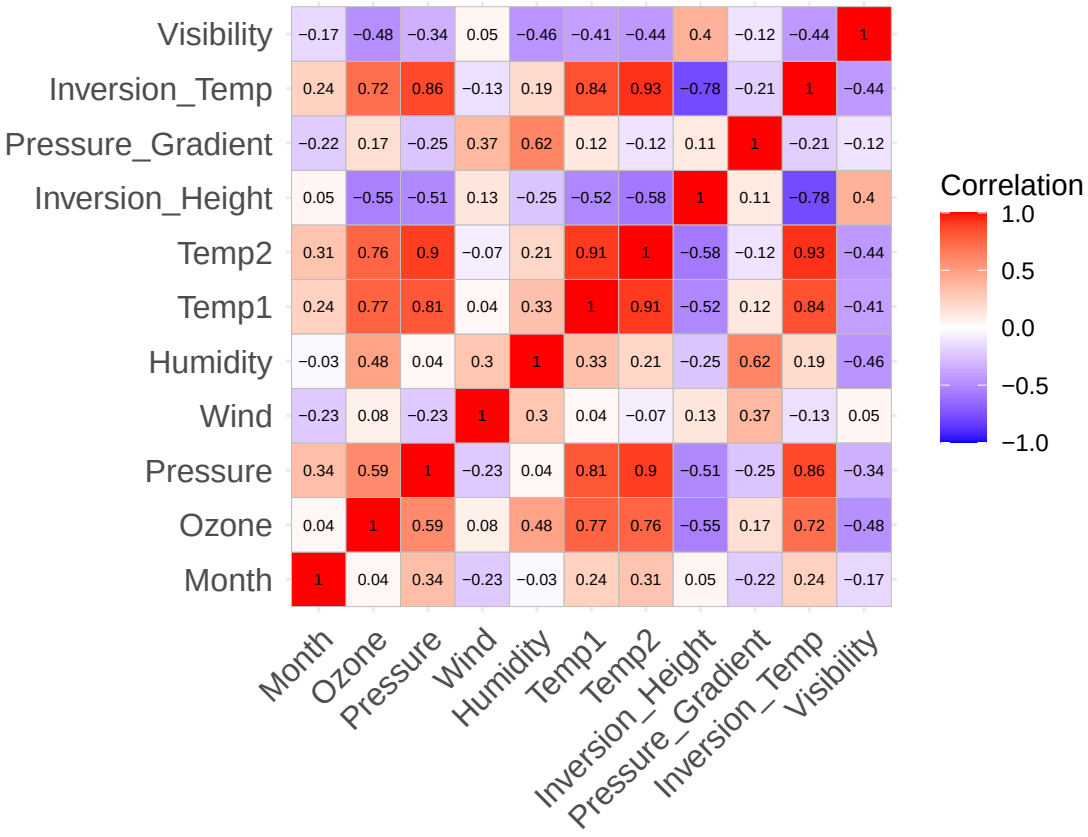
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
library(ggcorrplot)
corr_matrix <- cor(Ozone)

ggcorrplot(corr_matrix, lab = TRUE, lab_size = 2, colors = c("blue", "white", "red"), legend.title = "C
```



Task 1 Fit a linear model to the data.

Fit the full lm model with all variables

```
full_lm <- lm(Ozone ~ ., data = Ozone)
summary(full_lm)
```

```
##
## Call:
## lm(formula = Ozone ~ ., data = Ozone)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1296  -2.9738  -0.4418   2.6463  12.9798
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       55.5107446 37.3225049   1.487 0.138569
## Month             -0.3441411  0.1002204  -3.434 0.000729 ***
## Pressure          -0.0133351  0.0070585  -1.889 0.060369 .
## Wind              -0.0838465  0.1725545  -0.486 0.627583
## Humidity           0.0894286  0.0232652   3.844 0.000165 ***
## Temp1              0.1432746  0.0674711   2.123 0.034993 *
## Temp2              0.5516167  0.1216077   4.536 1.01e-05 ***
## Inversion_Height  -0.0006414  0.0003981  -1.611 0.108841
## Pressure_Gradient -0.0015969  0.0145076  -0.110 0.912463
## Inversion_Temp    -0.1263016  0.1163781  -1.085 0.279163
```

```
## Visibility        -0.0049013  0.0047850  -1.024 0.306977
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.361 on 192 degrees of freedom
## Multiple R-squared:  0.7305, Adjusted R-squared:  0.7164
## F-statistic: 52.03 on 10 and 192 DF,  p-value: < 2.2e-16
```

Assessing Multicollinearity Using Variance Inflation Factor (VIF)

```
# CAR stands for Companion to Applied Regression
library(car)
```

```
## Loading required package: carData
```

```
vif_values <- vif(full_lm)
print(vif_values)
```

```
##             Month          Pressure              Wind          Humidity
##          1.378504          6.759172          1.401590          2.498207
##             Temp1             Temp2  Inversion_Height Pressure_Gradient
##          9.756762         21.654794          5.823035          2.947910
##    Inversion_Temp        Visibility
##         28.694686          1.602020
```

Perform backward stepwise selection

```
backward_full_lm <- step(full_lm, direction = "backward")
```

```
## Start:  AIC=608.65
## Ozone ~ Month + Pressure + Wind + Humidity + Temp1 + Temp2 +
##     Inversion_Height + Pressure_Gradient + Inversion_Temp + Visibility
##
##                      Df Sum of Sq    RSS    AIC
## - Pressure_Gradient   1      0.23 3652.4 606.66
## - Wind                1      4.49 3656.7 606.89
## - Visibility          1     19.96 3672.2 607.75
## - Inversion_Temp      1     22.40 3674.6 607.89
## <none>                            3652.2 608.65
## - Inversion_Height    1     49.36 3701.6 609.37
## - Pressure            1     67.89 3720.1 610.38
## - Temp1               1     85.77 3738.0 611.36
## - Month               1    224.29 3876.5 618.74
## - Humidity            1    281.05 3933.3 621.70
## - Temp2               1    391.39 4043.6 627.31
##
## Step:  AIC=606.66
## Ozone ~ Month + Pressure + Wind + Humidity + Temp1 + Temp2 +
##     Inversion_Height + Inversion_Temp + Visibility
##
##                     Df Sum of Sq    RSS    AIC
## - Wind               1      4.61 3657.0 604.91
## - Visibility         1     20.08 3672.5 605.77
## - Inversion_Temp     1     23.46 3675.9 605.96
## <none>                           3652.4 606.66
```

```
## - Inversion_Height  1     49.49 3701.9 607.39
## - Pressure          1     67.66 3720.1 608.38
## - Temp1             1    107.08 3759.5 610.52
## - Month             1    226.34 3878.8 616.86
## - Humidity          1    380.59 4033.0 624.78
## - Temp2             1    392.50 4044.9 625.38
##
## Step:  AIC=604.91
## Ozone ~ Month + Pressure + Humidity + Temp1 + Temp2 + Inversion_Height +
##     Inversion_Temp + Visibility
##
##                    Df Sum of Sq    RSS    AIC
## - Visibility        1     22.73 3679.8 604.17
## - Inversion_Temp    1     23.44 3680.5 604.21
## <none>                          3657.0 604.91
## - Inversion_Height  1     53.45 3710.5 605.86
## - Pressure          1     63.13 3720.2 606.39
## - Temp1             1    103.06 3760.1 608.56
## - Month             1    222.43 3879.5 614.90
## - Humidity          1    380.35 4037.4 623.00
## - Temp2             1    387.89 4044.9 623.38
##
## Step:  AIC=604.17
## Ozone ~ Month + Pressure + Humidity + Temp1 + Temp2 + Inversion_Height +
##     Inversion_Temp
##
##                    Df Sum of Sq    RSS    AIC
## - Inversion_Temp    1     25.24 3705.0 603.56
## <none>                          3679.8 604.17
## - Pressure          1     61.97 3741.7 605.56
## - Inversion_Height  1     63.14 3742.9 605.63
## - Temp1             1     94.05 3773.8 607.30
## - Month             1    208.48 3888.2 613.36
## - Temp2             1    422.48 4102.2 624.23
## - Humidity          1    534.97 4214.7 629.73
##
## Step:  AIC=603.56
## Ozone ~ Month + Pressure + Humidity + Temp1 + Temp2 + Inversion_Height
##
##                    Df Sum of Sq    RSS    AIC
## <none>                          3705.0 603.56
## - Inversion_Height  1     45.88 3750.9 604.06
## - Pressure          1     80.61 3785.6 605.93
## - Temp1             1     87.16 3792.2 606.28
## - Month             1    229.65 3934.7 613.77
## - Temp2             1    516.30 4221.3 628.04
## - Humidity          1    601.20 4306.2 632.09
```

Fit the model with only significant variables

```
reduced_lm <- lm(Ozone ~ Inversion_Height + Pressure + Temp2 + Month + Humidity, data = Ozone)
summary(reduced_lm)
```

```
##
```

```
## Call:
## lm(formula = Ozone ~ Inversion_Height + Pressure + Temp2 + Month +
##     Humidity, data = Ozone)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4821  -3.0735  -0.0492   3.1372  13.0067
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      46.0035941 35.2678799   1.304 0.193617
## Inversion_Height -0.0002686  0.0002153  -1.247 0.213775
## Pressure         -0.0125756  0.0066372  -1.895 0.059597 .
## Temp2             0.6065083  0.0674621   8.990  < 2e-16 ***
## Month            -0.3637459  0.0954780  -3.810 0.000186 ***
## Humidity          0.1112523  0.0163736   6.795 1.26e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.387 on 197 degrees of freedom
## Multiple R-squared:  0.7201, Adjusted R-squared:  0.713
## F-statistic: 101.4 on 5 and 197 DF,  p-value: < 2.2e-16
```

Task 2 Fit a GLM with Gaussian Family

Fit the full GLM with Gaussian Family

```
full_glm_gaussian <- glm(Ozone ~ ., family = gaussian() , data = Ozone)
summary(full_glm_gaussian)
```

```
##
## Call:
## glm(formula = Ozone ~ ., family = gaussian(), data = Ozone)
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       55.5107446 37.3225049   1.487 0.138569
## Month             -0.3441411  0.1002204  -3.434 0.000729 ***
## Pressure          -0.0133351  0.0070585  -1.889 0.060369 .
## Wind              -0.0838465  0.1725545  -0.486 0.627583
## Humidity           0.0894286  0.0232652   3.844 0.000165 ***
## Temp1              0.1432746  0.0674711   2.123 0.034993 *
## Temp2              0.5516167  0.1216077   4.536 1.01e-05 ***
## Inversion_Height  -0.0006414  0.0003981  -1.611 0.108841
## Pressure_Gradient -0.0015969  0.0145076  -0.110 0.912463
## Inversion_Temp    -0.1263016  0.1163781  -1.085 0.279163
## Visibility        -0.0049013  0.0047850  -1.024 0.306977
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 19.02186)
##
##     Null deviance: 13549.5  on 202  degrees of freedom
## Residual deviance:  3652.2  on 192  degrees of freedom
```

```
## AIC: 1186.7
##
## Number of Fisher Scoring iterations: 2
```

Assessing Multicollinearity Using Variance Inflation Factor (VIF)

```
vif_values <- vif(full_glm_gaussian)
print(vif_values)
```

```
##             Month          Pressure              Wind          Humidity
##          1.378504          6.759172          1.401590          2.498207
##             Temp1             Temp2  Inversion_Height Pressure_Gradient
##          9.756762         21.654794          5.823035          2.947910
##    Inversion_Temp        Visibility
##         28.694686          1.602020
```

Perform backward stepwise selection

```
backward_full_glm_gaussian <- step(full_glm_gaussian, direction = "backward")
```

```
## Start:  AIC=1186.73
## Ozone ~ Month + Pressure + Wind + Humidity + Temp1 + Temp2 +
##     Inversion_Height + Pressure_Gradient + Inversion_Temp + Visibility
##
##                      Df Deviance    AIC
## - Pressure_Gradient   1   3652.4 1184.8
## - Wind                1   3656.7 1185.0
## - Visibility          1   3672.2 1185.8
## - Inversion_Temp      1   3674.6 1186.0
## <none>                    3652.2 1186.7
## - Inversion_Height    1   3701.6 1187.5
## - Pressure            1   3720.1 1188.5
## - Temp1               1   3738.0 1189.5
## - Month               1   3876.5 1196.8
## - Humidity            1   3933.3 1199.8
## - Temp2               1   4043.6 1205.4
##
## Step:  AIC=1184.75
## Ozone ~ Month + Pressure + Wind + Humidity + Temp1 + Temp2 +
##     Inversion_Height + Inversion_Temp + Visibility
##
##                     Df Deviance    AIC
## - Wind               1   3657.0 1183.0
## - Visibility         1   3672.5 1183.9
## - Inversion_Temp     1   3675.9 1184.0
## <none>                   3652.4 1184.8
## - Inversion_Height   1   3701.9 1185.5
## - Pressure           1   3720.1 1186.5
## - Temp1              1   3759.5 1188.6
## - Month              1   3878.8 1195.0
## - Humidity           1   4033.0 1202.9
## - Temp2              1   4044.9 1203.5
##
## Step:  AIC=1183
```

```
## Ozone ~ Month + Pressure + Humidity + Temp1 + Temp2 + Inversion_Height +
##     Inversion_Temp + Visibility
##
##                     Df Deviance    AIC
## - Visibility         1   3679.8 1182.3
## - Inversion_Temp     1   3680.5 1182.3
## <none>                   3657.0 1183.0
## - Inversion_Height   1   3710.5 1184.0
## - Pressure           1   3720.2 1184.5
## - Temp1              1   3760.1 1186.6
## - Month              1   3879.5 1193.0
## - Humidity           1   4037.4 1201.1
## - Temp2              1   4044.9 1201.5
##
## Step:  AIC=1182.26
## Ozone ~ Month + Pressure + Humidity + Temp1 + Temp2 + Inversion_Height +
##     Inversion_Temp
##
##                     Df Deviance    AIC
## - Inversion_Temp     1   3705.0 1181.7
## <none>                   3679.8 1182.3
## - Pressure           1   3741.7 1183.7
## - Inversion_Height   1   3742.9 1183.7
## - Temp1              1   3773.8 1185.4
## - Month              1   3888.2 1191.5
## - Temp2              1   4102.2 1202.3
## - Humidity           1   4214.7 1207.8
##
## Step:  AIC=1181.65
## Ozone ~ Month + Pressure + Humidity + Temp1 + Temp2 + Inversion_Height
##
##                     Df Deviance    AIC
## <none>                   3705.0 1181.7
## - Inversion_Height   1   3750.9 1182.2
## - Pressure           1   3785.6 1184.0
## - Temp1              1   3792.2 1184.4
## - Month              1   3934.7 1191.9
## - Temp2              1   4221.3 1206.1
## - Humidity           1   4306.2 1210.2
```

Fit the model with only significant variables

```
reduced_glm_gaussian <- glm(Ozone ~ Inversion_Height + Pressure + Temp2 + Month + Humidity, data = Ozone
summary(reduced_glm_gaussian)
```

```
##
## Call:
## glm(formula = Ozone ~ Inversion_Height + Pressure + Temp2 + Month +
##     Humidity, data = Ozone)
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      46.0035941 35.2678799   1.304 0.193617
## Inversion_Height -0.0002686  0.0002153  -1.247 0.213775
```

```
## Pressure         -0.0125756  0.0066372  -1.895 0.059597 .
## Temp2             0.6065083  0.0674621   8.990  < 2e-16 ***
## Month            -0.3637459  0.0954780  -3.810 0.000186 ***
## Humidity          0.1112523  0.0163736   6.795 1.26e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 19.24958)
##
##     Null deviance: 13549.5  on 202  degrees of freedom
## Residual deviance:  3792.2  on 197  degrees of freedom
## AIC: 1184.4
##
## Number of Fisher Scoring iterations: 2
```

Intermediate Comaparison

Compare AIC of full_lm, full_glm_gaussian, reduced_lm. reduced_glm_gaussian

```
AIC(full_lm, full_glm_gaussian, reduced_lm ,reduced_glm_gaussian)
```

```
##                       df      AIC
## full_lm               12 1186.734
## full_glm_gaussian     12 1186.734
## reduced_lm             7 1184.369
## reduced_glm_gaussian   7 1184.369
```

Task 3 Fit a GLM with Gamma Family

Fit the full GLM with Gamma Family

```
full_glm_gamma <- glm(Ozone ~., data = Ozone)
summary(full_glm_gamma)
```

```
##
## Call:
## glm(formula = Ozone ~ ., data = Ozone)
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       55.5107446 37.3225049   1.487 0.138569
## Month             -0.3441411  0.1002204  -3.434 0.000729 ***
## Pressure          -0.0133351  0.0070585  -1.889 0.060369 .
## Wind              -0.0838465  0.1725545  -0.486 0.627583
## Humidity           0.0894286  0.0232652   3.844 0.000165 ***
## Temp1              0.1432746  0.0674711   2.123 0.034993 *
## Temp2              0.5516167  0.1216077   4.536 1.01e-05 ***
## Inversion_Height  -0.0006414  0.0003981  -1.611 0.108841
## Pressure_Gradient -0.0015969  0.0145076  -0.110 0.912463
## Inversion_Temp    -0.1263016  0.1163781  -1.085 0.279163
## Visibility        -0.0049013  0.0047850  -1.024 0.306977
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 19.02186)
```

```
##
##     Null deviance: 13549.5  on 202  degrees of freedom
## Residual deviance:  3652.2  on 192  degrees of freedom
## AIC: 1186.7
##
## Number of Fisher Scoring iterations: 2
```

Assessing Multicollinearity Using Variance Inflation Factor (VIF)

```
vif_values <- vif(full_glm_gamma)
print(vif_values)
```

```
##             Month          Pressure              Wind          Humidity
##          1.378504          6.759172          1.401590          2.498207
##             Temp1             Temp2  Inversion_Height Pressure_Gradient
##          9.756762         21.654794          5.823035          2.947910
##    Inversion_Temp        Visibility
##         28.694686          1.602020
```

Perform backward stepwise selection

```
backward_full_glm_gamma <- step(full_glm_gamma, direction = "backward")
```

```
## Start:  AIC=1186.73
## Ozone ~ Month + Pressure + Wind + Humidity + Temp1 + Temp2 +
##     Inversion_Height + Pressure_Gradient + Inversion_Temp + Visibility
##
##                       Df Deviance    AIC
## - Pressure_Gradient    1   3652.4 1184.8
## - Wind                 1   3656.7 1185.0
## - Visibility           1   3672.2 1185.8
## - Inversion_Temp       1   3674.6 1186.0
## <none>                     3652.2 1186.7
## - Inversion_Height     1   3701.6 1187.5
## - Pressure             1   3720.1 1188.5
## - Temp1                1   3738.0 1189.5
## - Month                1   3876.5 1196.8
## - Humidity             1   3933.3 1199.8
## - Temp2                1   4043.6 1205.4
##
## Step:  AIC=1184.75
## Ozone ~ Month + Pressure + Wind + Humidity + Temp1 + Temp2 +
##     Inversion_Height + Inversion_Temp + Visibility
##
##                       Df Deviance    AIC
## - Wind                 1   3657.0 1183.0
## - Visibility           1   3672.5 1183.9
## - Inversion_Temp       1   3675.9 1184.0
## <none>                     3652.4 1184.8
## - Inversion_Height     1   3701.9 1185.5
## - Pressure             1   3720.1 1186.5
## - Temp1                1   3759.5 1188.6
## - Month                1   3878.8 1195.0
## - Humidity             1   4033.0 1202.9
```

```
## - Temp2             1   4044.9 1203.5
##
## Step:  AIC=1183
## Ozone ~ Month + Pressure + Humidity + Temp1 + Temp2 + Inversion_Height +
##     Inversion_Temp + Visibility
##
##                    Df Deviance    AIC
## - Visibility        1   3679.8 1182.3
## - Inversion_Temp    1   3680.5 1182.3
## <none>                  3657.0 1183.0
## - Inversion_Height  1   3710.5 1184.0
## - Pressure          1   3720.2 1184.5
## - Temp1             1   3760.1 1186.6
## - Month             1   3879.5 1193.0
## - Humidity          1   4037.4 1201.1
## - Temp2             1   4044.9 1201.5
##
## Step:  AIC=1182.26
## Ozone ~ Month + Pressure + Humidity + Temp1 + Temp2 + Inversion_Height +
##     Inversion_Temp
##
##                    Df Deviance    AIC
## - Inversion_Temp    1   3705.0 1181.7
## <none>                  3679.8 1182.3
## - Pressure          1   3741.7 1183.7
## - Inversion_Height  1   3742.9 1183.7
## - Temp1             1   3773.8 1185.4
## - Month             1   3888.2 1191.5
## - Temp2             1   4102.2 1202.3
## - Humidity          1   4214.7 1207.8
##
## Step:  AIC=1181.65
## Ozone ~ Month + Pressure + Humidity + Temp1 + Temp2 + Inversion_Height
##
##                    Df Deviance    AIC
## <none>                  3705.0 1181.7
## - Inversion_Height  1   3750.9 1182.2
## - Pressure          1   3785.6 1184.0
## - Temp1             1   3792.2 1184.4
## - Month             1   3934.7 1191.9
## - Temp2             1   4221.3 1206.1
## - Humidity          1   4306.2 1210.2
```

Fit the model with only significant variables

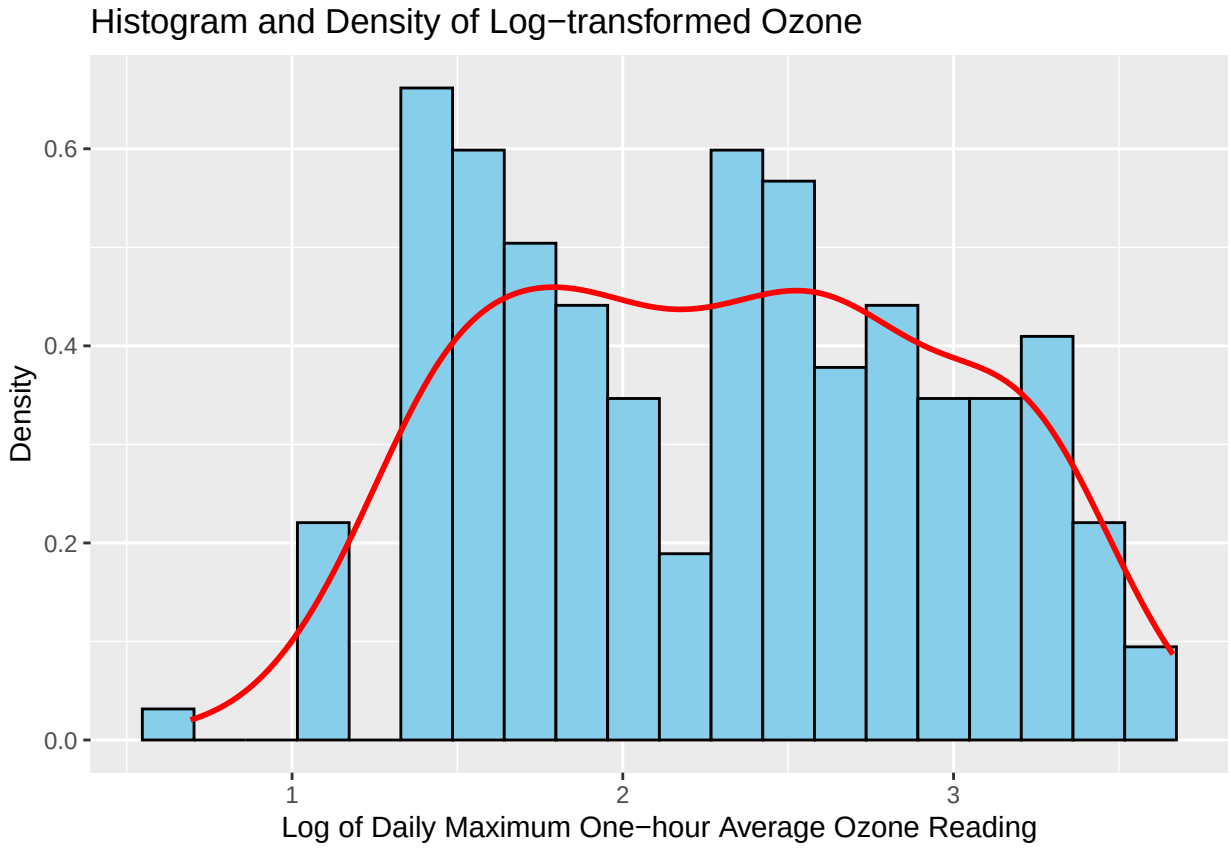
```
reduced_glm_gamma <- glm(Ozone ~ Inversion_Height + Pressure + Temp2 + Month + Humidity,
                         family = Gamma(), data = Ozone)
summary(reduced_glm_gamma)
```

```
##
## Call:
## glm(formula = Ozone ~ Inversion_Height + Pressure + Temp2 + Month +
##     Humidity, family = Gamma(), data = Ozone)
##
```

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.903e-01  3.651e-01   1.617  0.10751
## Inversion_Height  9.315e-06  2.334e-06   3.991 9.27e-05 ***
## Pressure         -5.675e-05  6.732e-05  -0.843  0.40022
## Temp2            -2.359e-03  5.139e-04  -4.591 7.86e-06 ***
## Month             2.902e-03  9.889e-04   2.934  0.00374 **
## Humidity         -9.522e-04  1.580e-04  -6.026 8.12e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.1772186)
##
##     Null deviance: 110.611  on 202  degrees of freedom
## Residual deviance:  38.363  on 197  degrees of freedom
## AIC: 1134.8
##
## Number of Fisher Scoring iterations: 5
```

Task 4 Log(Ozone) + GLM Gamma

```r
# Shift the Ozone data by adding a small constant to avoid zeros
Ozone$log_Ozone <- log(Ozone$Ozone + 1)
# Plot the histogram and density of the log-transformed Ozone
ggplot(Ozone, aes(x = log_Ozone)) +
  geom_histogram(aes(y = ..density..), bins = 20, fill = "skyblue", color = "black") +
  geom_density(color = "red", size = 1) +
  labs(title = "Histogram and Density of Log-transformed Ozone",
       x = "Log of Daily Maximum One-hour Average Ozone Reading",
       y = "Density")
```



Fit the full GLM with the Gamma family on the log-transformed Ozone data

```
full_logozone_glm_gamma <- glm(log_Ozone ~ . - Ozone, family = Gamma(link = "inverse"), data = Ozone)
summary(full_logozone_glm_gamma)
```

```
##
## Call:
## glm(formula = log_Ozone ~ . - Ozone, family = Gamma(link = "inverse"),
##     data = Ozone)
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.871e-01  7.347e-01   1.071 0.285376
## Month               7.802e-03  1.931e-03   4.041 7.68e-05 ***
## Pressure           -3.383e-06  1.378e-04  -0.025 0.980437
## Wind                3.491e-03  3.086e-03   1.131 0.259366
## Humidity           -1.439e-03  4.157e-04  -3.462 0.000662 ***
## Temp1              -3.540e-03  1.233e-03  -2.871 0.004549 **
## Temp2              -4.972e-03  2.074e-03  -2.398 0.017451 *
## Inversion_Height    1.727e-05  6.721e-06   2.570 0.010937 *
## Pressure_Gradient   5.310e-05  2.545e-04   0.209 0.834970
## Inversion_Temp      2.498e-03  1.967e-03   1.270 0.205473
## Visibility          8.267e-05  9.336e-05   0.885 0.377010
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.03048865)
```

```
##
##     Null deviance: 19.1153  on 202  degrees of freedom
## Residual deviance:  6.3128  on 192  degrees of freedom
## AIC: 215.17
##
## Number of Fisher Scoring iterations: 4
```

Assessing Multicollinearity Using Variance Inflation Factor (VIF)

```
vif_values <- vif(full_logozone_glm_gamma )
print(vif_values)
```

```
##             Month          Pressure              Wind          Humidity
##          1.353341          7.369998          1.254317          2.278026
##             Temp1             Temp2  Inversion_Height Pressure_Gradient
##         11.739192         22.068460          4.590780          2.670188
##    Inversion_Temp        Visibility
##         26.408728          1.706697
```

Perform backward stepwise selection

```
backward_full_logozone_glm_gamma <- step(full_logozone_glm_gamma, direction = "backward")
```

```
## Start:  AIC=215.17
## log_Ozone ~ (Month + Ozone + Pressure + Wind + Humidity + Temp1 +
##     Temp2 + Inversion_Height + Pressure_Gradient + Inversion_Temp +
##     Visibility) - Ozone
##
##                     Df Deviance    AIC
## - Pressure           1   6.3129 213.17
## - Pressure_Gradient  1   6.3142 213.21
## - Visibility         1   6.3369 213.96
## - Wind               1   6.3519 214.45
## - Inversion_Temp     1   6.3624 214.79
## <none>                   6.3128 215.17
## - Temp2              1   6.4897 218.97
## - Inversion_Height   1   6.5147 219.79
## - Temp1              1   6.5675 221.52
## - Humidity           1   6.6766 225.10
## - Month              1   6.8086 229.43
##
## Step:  AIC=213.17
## log_Ozone ~ Month + Wind + Humidity + Temp1 + Temp2 + Inversion_Height +
##     Pressure_Gradient + Inversion_Temp + Visibility
##
##                     Df Deviance    AIC
## - Pressure_Gradient  1   6.3142 211.21
## - Visibility         1   6.3370 211.96
## - Wind               1   6.3545 212.54
## - Inversion_Temp     1   6.3632 212.83
## <none>                   6.3129 213.17
## - Temp2              1   6.5096 217.65
## - Inversion_Height   1   6.5163 217.88
## - Temp1              1   6.5728 219.74
```

```
## - Humidity           1   6.6861 223.47
## - Month              1   6.8117 227.62
##
## Step:  AIC=211.21
## log_Ozone ~ Month + Wind + Humidity + Temp1 + Temp2 + Inversion_Height +
##     Inversion_Temp + Visibility
##
##                    Df Deviance    AIC
## - Visibility        1   6.3392 210.04
## - Wind              1   6.3564 210.61
## - Inversion_Temp    1   6.3685 211.01
## <none>                  6.3142 211.21
## - Temp2             1   6.5099 215.70
## - Inversion_Height  1   6.5189 216.00
## - Temp1             1   6.6323 219.75
## - Humidity          1   6.7835 224.77
## - Month             1   6.8176 225.90
##
## Step:  AIC=210.02
## log_Ozone ~ Month + Wind + Humidity + Temp1 + Temp2 + Inversion_Height +
##     Inversion_Temp
##
##                    Df Deviance    AIC
## - Wind              1   6.3855 209.55
## - Inversion_Temp    1   6.3920 209.77
## <none>                  6.3392 210.02
## - Temp2             1   6.5563 215.21
## - Inversion_Height  1   6.5594 215.31
## - Temp1             1   6.6421 218.05
## - Month             1   6.8265 224.16
## - Humidity          1   7.0215 230.62
##
## Step:  AIC=209.5
## log_Ozone ~ Month + Humidity + Temp1 + Temp2 + Inversion_Height +
##     Inversion_Temp
##
##                    Df Deviance    AIC
## - Inversion_Temp    1   6.4342 209.11
## <none>                  6.3855 209.50
## - Temp2             1   6.6045 214.73
## - Inversion_Height  1   6.6277 215.50
## - Temp1             1   6.6607 216.59
## - Month             1   6.8357 222.36
## - Humidity          1   7.0220 228.51
##
## Step:  AIC=209.05
## log_Ozone ~ Month + Humidity + Temp1 + Temp2 + Inversion_Height
##
##                    Df Deviance    AIC
## <none>                  6.4342 209.05
## - Temp2             1   6.6321 213.59
## - Inversion_Height  1   6.6958 215.70
## - Temp1             1   6.6991 215.81
## - Month             1   6.9426 223.85
```

```
## - Humidity          1   7.2092 232.66
```

Fit the model with only significant variables

```
reduced_logozone_glm_gamma <- glm(log_Ozone ~ Temp2 + Inversion_Height + Month + Humidity,
                          family = Gamma(link = "log"), data = Ozone)
summary(reduced_logozone_glm_gamma)
```

```
##
## Call:
## glm(formula = log_Ozone ~ Temp2 + Inversion_Height + Month +
##     Humidity, family = Gamma(link = "log"), data = Ozone)
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.480e-01  9.265e-02  -2.676  0.00806 **
## Temp2             1.744e-02  1.400e-03  12.455  < 2e-16 ***
## Inversion_Height -1.683e-05  8.425e-06  -1.998  0.04711 *
## Month            -2.013e-02  3.717e-03  -5.415 1.76e-07 ***
## Humidity          4.123e-03  6.025e-04   6.843 9.46e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.02956796)
##
##     Null deviance: 19.1153  on 202  degrees of freedom
## Residual deviance:  6.4312  on 198  degrees of freedom
## AIC: 206.96
##
## Number of Fisher Scoring iterations: 5
```

Task 5 Final Comaparison

Compare AIC of full_lm, full_glm_gaussian, full_glm_gamma, reduced_lm. reduced_glm_gaussian, reduced_glm_gamma

```
AIC(full_lm, full_glm_gaussian, full_glm_gamma, reduced_lm ,reduced_glm_gaussian, reduced_glm_gamma, re
```

```
##                             df       AIC
## full_lm                     12 1186.7343
## full_glm_gaussian           12 1186.7343
## full_glm_gamma              12 1186.7343
## reduced_lm                   7 1184.3689
## reduced_glm_gaussian         7 1184.3689
## reduced_glm_gamma            7 1134.7743
## reduced_logozone_glm_gamma   6  206.9598
```

```
library(Metrics)

# Predictions for each model
pred_full_lm <- predict(full_lm, Ozone)
pred_full_glm_gaussian <- predict(full_glm_gaussian, Ozone)
pred_full_glm_gamma <- predict(full_glm_gamma, Ozone)
pred_full_logozone_glm_gamma <- predict(full_logozone_glm_gamma, Ozone)
pred_reduced_lm <- predict(reduced_lm, Ozone)
```

```r
pred_reduced_glm_gaussian <- predict(reduced_glm_gaussian, Ozone)
pred_reduced_glm_gamma <- predict(reduced_glm_gamma, Ozone)
pred_reduced_logozone_glm_gamma <- predict(reduced_logozone_glm_gamma, Ozone)

# Actual values
actual_values <- Ozone$Ozone

# Calculate RMSE for each model
rmse_full_lm <- rmse(actual_values, pred_full_lm)
rmse_full_glm_gaussian <- rmse(actual_values, pred_full_glm_gaussian)
rmse_full_glm_gamma <- rmse(actual_values, pred_full_glm_gamma)
rmse_full_logozone_glm_gamma <- rmse(Ozone$log_Ozone, pred_full_logozone_glm_gamma)

rmse_reduced_lm <- rmse(actual_values, pred_reduced_lm)
rmse_reduced_glm_gaussian <- rmse(actual_values, pred_reduced_glm_gaussian)
rmse_reduced_glm_gamma <- rmse(actual_values, pred_reduced_glm_gamma)
rmse_reduced_logozone_glm_gamma <- rmse(Ozone$log_Ozone, pred_reduced_logozone_glm_gamma)

# Compare RMSE values
rmse_values <- data.frame(
  Model = c("Full LM", "Full GLM Gaussian", "Full GLM Gamma",
            "Full LogOzone GLM Gamma",  # New full model
            "Reduced LM", "Reduced GLM Gaussian", "Reduced GLM Gamma",
            "Reduced LogOzone GLM Gamma"),  # New reduced model
  RMSE = c(rmse_full_lm, rmse_full_glm_gaussian, rmse_full_glm_gamma,
           rmse_full_logozone_glm_gamma,  # New full model RMSE
           rmse_reduced_lm, rmse_reduced_glm_gaussian, rmse_reduced_glm_gamma,
           rmse_reduced_logozone_glm_gamma)  # New reduced model RMSE
)

# Display RMSE values for comparison
print(rmse_values)
```

```
##                       Model      RMSE
## 1                   Full LM  4.241594
## 2          Full GLM Gaussian  4.241594
## 3             Full GLM Gamma  4.241594
## 4    Full LogOzone GLM Gamma  1.988657
## 5                 Reduced LM  4.322109
## 6       Reduced GLM Gaussian  4.322109
## 7          Reduced GLM Gamma 13.933418
## 8 Reduced LogOzone GLM Gamma  1.571191
```

```r
library(Metrics)

# Predictions for each model
pred_full_lm <- predict(full_lm, Ozone)
pred_full_glm_gaussian <- predict(full_glm_gaussian, Ozone)
pred_full_glm_gamma <- predict(full_glm_gamma, Ozone)
pred_full_logozone_glm_gamma <- predict(full_logozone_glm_gamma, Ozone)

pred_reduced_lm <- predict(reduced_lm, Ozone)
pred_reduced_glm_gaussian <- predict(reduced_glm_gaussian, Ozone)
pred_reduced_glm_gamma <- predict(reduced_glm_gamma, Ozone)
```

```r
pred_reduced_logozone_glm_gamma <- predict(reduced_logozone_glm_gamma, Ozone)

# Actual values
actual_values <- Ozone$Ozone

# Calculate MAE for each model
mae_full_lm <- mae(actual_values, pred_full_lm)
mae_full_glm_gaussian <- mae(actual_values, pred_full_glm_gaussian)
mae_full_glm_gamma <- mae(actual_values, pred_full_glm_gamma)
mae_full_logozone_glm_gamma <- mae(Ozone$log_Ozone, pred_full_logozone_glm_gamma)

mae_reduced_lm <- mae(actual_values, pred_reduced_lm)
mae_reduced_glm_gaussian <- mae(actual_values, pred_reduced_glm_gaussian)
mae_reduced_glm_gamma <- mae(actual_values, pred_reduced_glm_gamma)
mae_reduced_logozone_glm_gamma <- mae(Ozone$log_Ozone, pred_reduced_logozone_glm_gamma)

# Store and compare the MAE values
mae_values <- data.frame(
  Model = c("Full LM", "Full GLM Gaussian", "Full GLM Gamma",
            "Full LogOzone GLM Gamma",  # Full LogOzone GLM Gamma added
            "Reduced LM", "Reduced GLM Gaussian", "Reduced GLM Gamma",
            "Reduced LogOzone GLM Gamma"),  # Reduced LogOzone GLM Gamma added
  MAE = c(mae_full_lm, mae_full_glm_gaussian, mae_full_glm_gamma,
          mae_full_logozone_glm_gamma,  # MAE for full LogOzone GLM Gamma
          mae_reduced_lm, mae_reduced_glm_gaussian, mae_reduced_glm_gamma,
          mae_reduced_logozone_glm_gamma)  # MAE for reduced LogOzone GLM Gamma
)

# Display MAE values for comparison
print(mae_values)
```

```
##                         Model       MAE
## 1                     Full LM  3.416843
## 2           Full GLM Gaussian  3.416843
## 3              Full GLM Gamma  3.416843
## 4    Full LogOzone GLM Gamma  1.834642
## 5                  Reduced LM  3.543576
## 6        Reduced GLM Gaussian  3.543576
## 7           Reduced GLM Gamma 11.255267
## 8 Reduced LogOzone GLM Gamma  1.496788
```

Final Model Comparison and Conclusion

1. AIC (Akaike Information Criterion) Comparison

- Lower AIC values indicate a better balance between goodness-of-fit and model complexity.
- The **Reduced LogOzone GLM Gamma** model has the **lowest AIC (206.96)**, making it the best model based on model selection criteria.
- Other models (Full LM, Full GLM Gaussian, Full GLM Gamma) have similar AIC values of around **1186**, which are much higher, indicating worse performance compared to the reduced LogOzone GLM Gamma model.

2. RMSE (Root Mean Squared Error) Comparison

- RMSE measures prediction accuracy, with lower values being better.
- The **Reduced LogOzone GLM Gamma** model has the **lowest RMSE (1.571)**, followed closely by the **Full LogOzone GLM Gamma** with **RMSE = 1.988**.
- The original models (Full LM, Full GLM Gaussian, Full GLM Gamma) have much higher RMSE values of around **4.24**, indicating less accurate predictions compared to the LogOzone models.

3. MAE (Mean Absolute Error) Comparison

- MAE measures the average magnitude of prediction errors, with lower values indicating better performance.
- The **Reduced LogOzone GLM Gamma** model has the **lowest MAE (1.497)**, followed by the **Full LogOzone GLM Gamma** with **MAE = 1.834**.
- The Full LM, Full GLM Gaussian, and Full GLM Gamma models have higher MAE values of **3.416**, and the Reduced GLM Gamma model performs the worst with an MAE of **11.25**.

Conclusion

Based on all three metrics (AIC, RMSE, and MAE): - The **Reduced LogOzone GLM Gamma** model performs the best overall. It has the lowest AIC, RMSE, and MAE, indicating that it balances model complexity, prediction accuracy, and overall fit better than the other models. - The **Full LogOzone GLM Gamma** model also performs well, especially in terms of prediction accuracy, though its AIC is slightly higher compared to the reduced version.

Therefore, the **Reduced LogOzone GLM Gamma** model is the best choice for predicting the daily maximum one-hour average ozone reading in this case.