Ozone concentration and meteorology in the LA Basin,1976 - A Regression Study

Arkajyoti Bhattacharjee
Vishweshwar Tyagi
Saurab Jain
Apoorva Singh

Indian Institute of Technology, Kanpur

About the project

Aim of the Study

- understand the relationship between **Ozone concentration** and meteorological variables like **temperature**, **pressure**, **humidity**, etc.
- develop **parametric** and **non-parametric** models to be able to **predict** ozone concentration based on given values of the meteorological variables.

What we have done

- fitted various regression models while **detecting** and taking **remedial measures** for the problems of **multi-collinearity**, **heteroscedasticity** and **auto-correlation** of **errors**.
- compared the **predictive power** of the models developed in the process by compairing the Root Mean Square Error(**RMSE**) of the model.

The Ozone Dataset and Exploratory Analysis

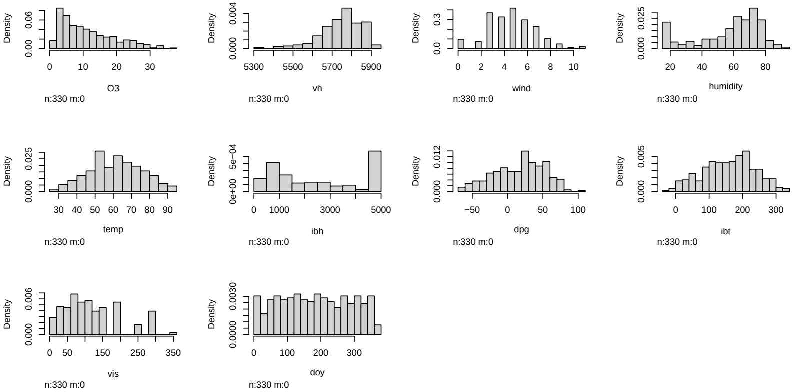
Data description

- **Ozone in Los Angeles Basin in 1976** dataset.
  - historical time-series data.
  - **330** observations and **10** variables.
- variables associated with this dataset -
  - **O3:** Ozone conc., ppm, at Sandbug AFB.
  - **vh:** a numeric vector
  - **wind:** wind speed
  - **humidity:** a numeric vector
  - **temp:** temperature
  - **ibh:** inversion base height
  - **dpg:** Daggett pressure gradient
  - **ibt:** a numeric vector
  - **vis:** visibility
  - **doy:** day of the year
- Here, **O3** is the response variable and the remaining are potential regressors.

Data Summary

```
##        O3              vh            wind           humidity    
##  Min.   : 1.00   Min.   :5320   Min.   : 0.000   Min.   :19.00  
##  1st Qu.: 5.00   1st Qu.:5690   1st Qu.: 3.000   1st Qu.:47.00  
##  Median :10.00   Median :5760   Median : 5.000   Median :64.00  
##  Mean   :11.78   Mean   :5750   Mean   : 4.848   Mean   :58.13  
##  3rd Qu.:17.00   3rd Qu.:5830   3rd Qu.: 6.000   3rd Qu.:73.00  
##  Max.   :38.00   Max.   :5950   Max.   :11.000   Max.   :93.00  
##       temp            ibh              dpg              ibt       
##  Min.   :25.00   Min.   : 111.0   Min.   :-69.00   Min.   :-25.0  
##  1st Qu.:51.00   1st Qu.: 877.5   1st Qu.: -9.00   1st Qu.:107.0  
##  Median :62.00   Median :2112.5   Median : 24.00   Median :167.5  
##  Mean   :61.75   Mean   :2572.9   Mean   : 17.37   Mean   :161.2  
##  3rd Qu.:72.00   3rd Qu.:5000.0   3rd Qu.: 44.75   3rd Qu.:214.0  
##  Max.   :93.00   Max.   :5000.0   Max.   :107.00   Max.   :332.0  
##       vis             doy         
##  Min.   :  0.0   Min.   :  1.00  
##  1st Qu.: 70.0   1st Qu.: 96.25  
##  Median :120.0   Median :182.50  
##  Mean   :124.5   Mean   :183.88  
##  3rd Qu.:150.0   3rd Qu.:273.75  
##  Max.   :350.0   Max.   :365.00  
```

Histograms of the Variables



Parametric Model Setup : Model Assumptions

- we first fit a multiple linear regression model to the data, with **O3** as the response and all other variables as regressors.
- The model is given by :

$$O_3 = \beta_0 + \beta_1 vh + \beta_2 humidity + \beta_3 wind + \beta_4 temp + \beta_5 dpg + \beta_6 ibt + \beta_7 ibh + \beta_8 doy + \beta_9 vis + \epsilon$$

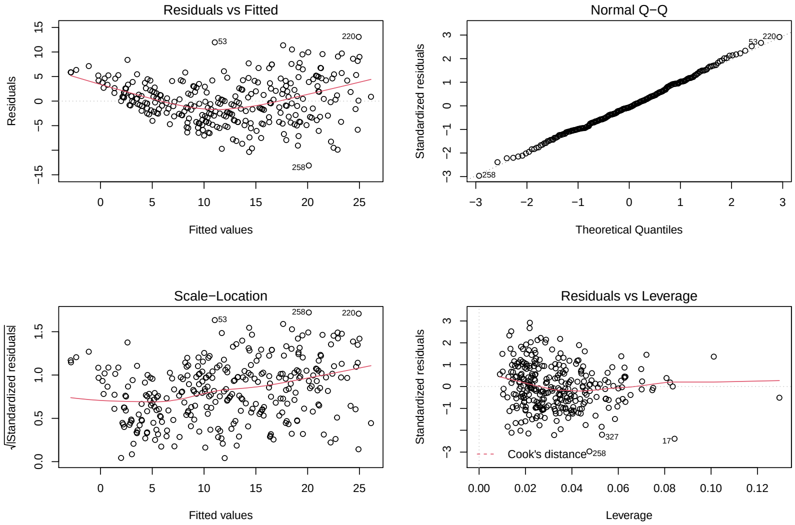
- assume a Gauss-Markov set-up i.e. we make the following assumptions:

1. $E(\epsilon) = 0$
2. $var(\epsilon) = \sigma^2 I$ i.e.
   2.1. $var(\epsilon_i) = \sigma^2 \ \forall \ i$
   2.2. $cov(\epsilon_i, \epsilon_j) = 0 \ \forall \ i \neq j$

- for testing purposes, we assume

3. $\epsilon \sim N(0, \sigma^2 I)$

Model 0 and Basic Diagnostic Plots



Summary of Model 0

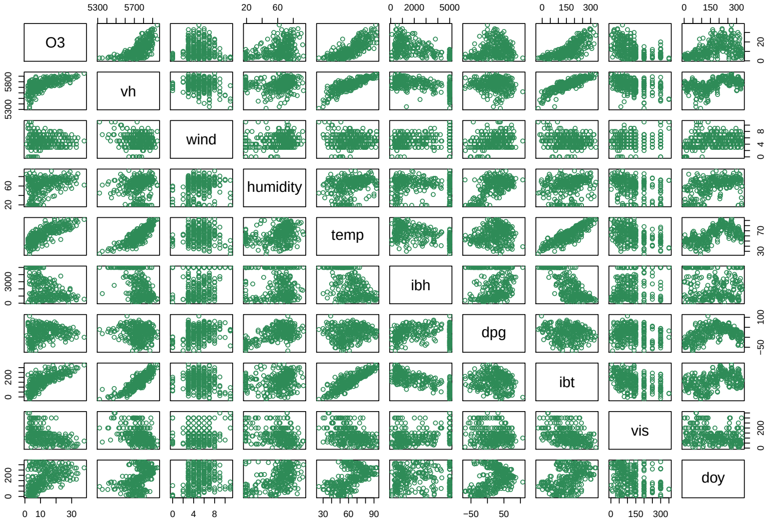
```
##
## Call:
## lm(formula = O3 ~ ., data = ozone[1:300, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1115  -2.9906  -0.2988   2.9341  13.0716
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.5006544 32.4565235   0.755 0.450936
## vh          -0.0062400  0.0059171  -1.055 0.292495
## wind         0.0328400  0.1491718   0.220 0.825910
## humidity     0.0771142  0.0213435   3.613 0.000357 ***
## temp         0.2647941  0.0520989   5.083 6.69e-07 ***
## ibh         -0.0004993  0.0003108  -1.607 0.109232
## dpg          0.0009924  0.0119021   0.083 0.933604
## ibt          0.0294090  0.0144697   2.032 0.043018 *
## vis         -0.0060750  0.0039846  -1.525 0.128450
## doy         -0.0023407  0.0041495  -0.564 0.573123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.53 on 290 degrees of freedom
## Multiple R-squared:  0.6986, Adjusted R-squared:  0.6892
## F-statistic: 74.68 on 9 and 290 DF,  p-value: < 2.2e-16
```

Remarks based on Graphs and Summary of Model 0

- based on the graphs, we observe -
  - There is curvature in the **residual vs fitted plot** indicating a **non-linear** relationship in the data-set.
  - There is **heteroscedasticity** in the data as the residuals do not form a constant band.
  - The **normal Q-Q** plot shows a fairly straight line, indicating the errors are more-or-less **normally distributed**.
  - 17, 53, 258 and 220^{th} observations may need special attention.
- based on the summary of the fitted model, we observe -
  - The **Multiple R-squared** of the model is: **0.6986** and the **Adjusted R-squared** is: **0.6892**.
  - Since the errors seem to follow normal distribution based on **Q-Q** plot, so taking level of significance to be 0.01, only **humidity** and **temperature** seem to be *statistically significant* based on their p-values.

Multicollinearity

Scatterplot Matrix



Remarks

Based on the **scatterplot matrix**, we observe -

- **vh** and **temp** seem to be almost perfectly **positively correlated**
- **temp** and **ibt** seem to be almost perfectly **positively correlated**
- As expected from the above two points, **vh** and **ibt** seem to be almost perfectly **positively correlated**
- **dpg** and **doy** have a somewhat quadratic relationship
- **temp** and **doy** have a somewhat quadratic relationship

Eigen-Decomposition Proportion(EDP)

```
##
## Call:
## eigprop(mod = lm(O3 ~ . - 1, data = ozone[1:300, ]))
##
##   Eigenvalues      CI     vh   wind humidity   temp    ibh    dpg    ibt    vis
## 1      7.1759  1.0000 0.0002 0.0019   0.0007 0.0001 0.0010 0.0022 0.0002 0.0027
## 2      0.7448  3.1041 0.0003 0.0002   0.0010 0.0000 0.0075 0.2862 0.0000 0.0421
## 3      0.6113  3.4261 0.0001 0.0016   0.0005 0.0006 0.0426 0.0645 0.0060 0.0380
## 4      0.1974  6.0295 0.0000 0.0011   0.0003 0.0000 0.1223 0.0782 0.0013 0.4798
## 5      0.1106  8.0540 0.0080 0.0562   0.0222 0.0010 0.0047 0.0005 0.0014 0.2589
## 6      0.0991  8.5073 0.0032 0.8726   0.0036 0.0026 0.0638 0.0387 0.0038 0.0028
## 7      0.0474 12.3076 0.0009 0.0390   0.5720 0.0104 0.0609 0.2352 0.0360 0.0244
## 8      0.0092 27.9955 0.8411 0.0268   0.3906 0.0002 0.4512 0.0221 0.2067 0.1512
## 9      0.0043 40.7511 0.1462 0.0006   0.0091 0.9850 0.2460 0.2724 0.7447 0.0000
##      doy
## 1 0.0018
## 2 0.0032
## 3 0.0097
## 4 0.0797
## 5 0.6228
## 6 0.0125
## 7 0.0165
## 8 0.2451
## 9 0.0086
##
## ===============================
## Row 6==> wind, proportion 0.872600 >= 0.50
## Row 7==> humidity, proportion 0.572021 >= 0.50
## Row 9==> temp, proportion 0.985027 >= 0.50
## Row 9==> ibt, proportion 0.744695 >= 0.50
## Row 5==> doy, proportion 0.622836 >= 0.50
```

Variance Inflation Factors(VIFs) and Remarks

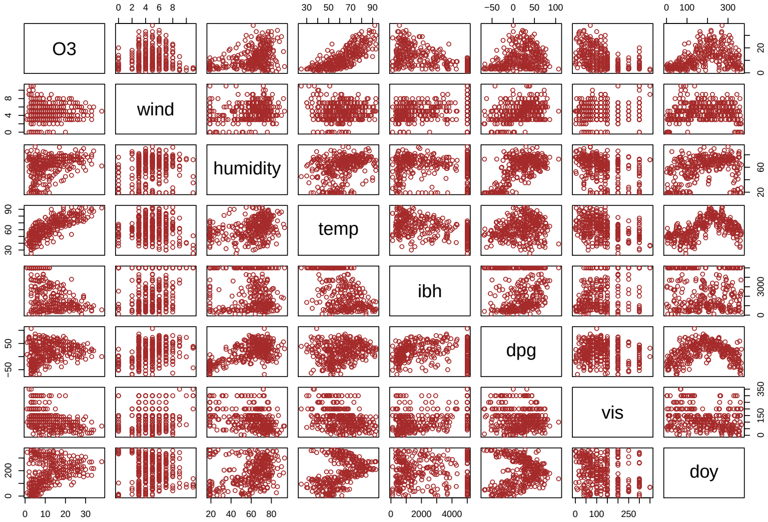
```
##        vh      wind  humidity      temp       ibh       dpg       ibt       vis
##  5.884904  1.282581  2.445097  8.624229  4.492747  2.465877 18.457599  1.426169
##       doy
##  2.266763
```

- **wind**, **temp**, **humidity**, **ibt** and **doy** have variance decompositon proportion greater than 0.50.
- **vh**, **temp** and **ibt** have **VIFs**$>$**5.**

Variable Drop(Model A)

```
##
## Call:
## lm(formula = O3 ~ . - ibt - vh, data = ozone[1:300, ])
##
## Coefficients:
## (Intercept)         wind     humidity         temp          ibh          dpg
##  -9.4404825    0.0567674    0.0780854    0.3295249   -0.0009882   -0.0084556
##         vis          doy
##  -0.0065565   -0.0015451

##     wind humidity     temp      ibh      dpg      vis      doy
## 1.227943 2.402486 2.367630 1.730002 1.867278 1.392424 2.143054

## The R^2 value of lmodA is :  0.6942595
```

Scatterplot Matrix after Variable Drop

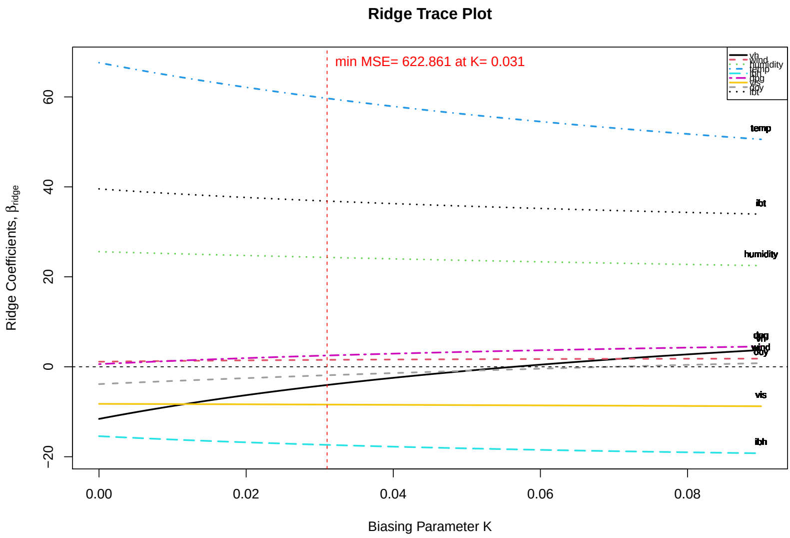


Remarks for Model A

We make the following observations based on the above scatterplot matrix -

- There is a quadratic relationship between **temp** and **doy**. This is expected as temperature increases in the middle of the year and is lower elsewhere.
- A similar relationship seems to exist between **dpg** and **doy**

Ridge Regression(Model B)



Model B: Summary and VIFs

```
##
## Call:
## lmridge.default(formula = O3 ~ vh + wind + humidity + temp +
##     ibh + dpg + vis + doy + ibt, data = ozone[1:300, ], K = 0.031)
##
##
## Coefficients: for Ridge parameter K= 0.031
##             Estimate Estimate (Sc) StdErr (Sc) t-value (Sc) Pr(>|t|)
## Intercept     3.5095    57822.5975  52104.9087       1.1097   0.2680
## vh           -0.0022       -4.0506      8.5375      -0.4745   0.6355
## wind          0.0448        1.5421      4.8737       0.3164   0.7519
## humidity      0.0733       24.3160      6.3794       3.8117   0.0002 ***
## temp          0.2337       59.6751      9.0778       6.5737   <2e-16 ***
## ibh          -0.0006      -17.3637      6.8103      -2.5496   0.0113 *
## dpg           0.0042        2.5090      6.0233       0.4166   0.6773
## vis          -0.0062       -8.4051      5.1317      -1.6379   0.1025
## doy          -0.0012       -1.8842      6.1923      -0.3043   0.7611
## ibt           0.0274       36.8441     10.7511       3.4270   0.0007 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Ridge Summary
##         R2     adj-R2   DF ridge           F        AIC         BIC
##    0.67910    0.67030    7.99421    74.77373  913.13516 2653.87872
## Ridge minimum MSE= 622.8613 at K= 0.031
## P-value for F-test ( 7.99421 , 291.2756 ) = 4.308328e-66
## ---------------------------------------------------------------

##               vh    wind humidity     temp     ibh    dpg      vis      doy      ibt
## k=0.031 3.55702 1.15917  1.98602 4.02152 2.26339 1.7705 1.28516 1.87128 5.64078
```

Principal Components Regression(Model C)

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.9906 1.4324 0.9824 0.80988 0.78021 0.60941 0.47795
## Proportion of Variance 0.4403 0.2280 0.1072 0.07288 0.06764 0.04126 0.02538
## Cumulative Proportion  0.4403 0.6683 0.7755 0.84840 0.91604 0.95730 0.98268
##                            PC8     PC9
## Standard deviation     0.34451 0.19278
## Proportion of Variance 0.01319 0.00413
## Cumulative Proportion  0.99587 1.00000
```

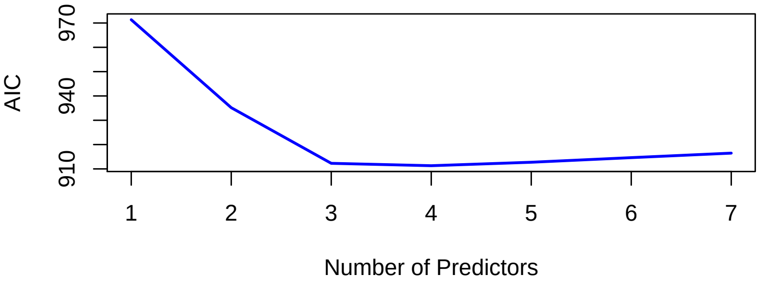
Model C: Summary, Regression Coefficients and VIFs

```
##
## Call:
## lm(formula = O3 ~ ., data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1115  -2.9906  -0.2988   2.9341  13.0716
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.20000    0.26152  46.651  < 2e-16 ***
## PC1          3.31921    0.13159  25.223  < 2e-16 ***
## PC2          0.12221    0.18287   0.668  0.50448
## PC3         -0.03486    0.26664  -0.131  0.89608
## PC4          0.97992    0.32345   3.030  0.00267 **
## PC5          0.51580    0.33575   1.536  0.12557
## PC6         -0.51336    0.42985  -1.194  0.23335
## PC7          1.17635    0.54808   2.146  0.03268 *
## PC8         -3.21286    0.76038  -4.225  3.2e-05 ***
## PC9         -0.08670    1.35881  -0.064  0.94917
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.53 on 290 degrees of freedom
## Multiple R-squared:  0.6986, Adjusted R-squared:  0.6892
## F-statistic: 74.68 on 9 and 290 DF,  p-value: < 2.2e-16

## The model parameter estimates are
##  -0.6701572 0.06531083 1.479936 3.909912 -0.892049 0.03430019 2.287367 -0.4769441 -0.2224769

## PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9
##   1   1   1   1   1   1   1   1   1
```

Variable Selection

Model A

```
##   (Intercept)  wind humidity temp   ibh   dpg   vis   doy
## 1        TRUE FALSE    FALSE TRUE FALSE FALSE FALSE FALSE
## 2        TRUE FALSE    FALSE TRUE  TRUE FALSE FALSE FALSE
## 3        TRUE FALSE     TRUE TRUE  TRUE FALSE FALSE FALSE
## 4        TRUE FALSE     TRUE TRUE  TRUE FALSE  TRUE FALSE
## 5        TRUE FALSE     TRUE TRUE  TRUE  TRUE  TRUE FALSE
## 6        TRUE  TRUE     TRUE TRUE  TRUE  TRUE  TRUE FALSE
## 7        TRUE  TRUE     TRUE TRUE  TRUE  TRUE  TRUE  TRUE

## Mallows Cp value for p in 1 to 7:  68.901 27.312 3.73 2.816 4.249 6.146 8

## Adjusted R^2 value for p in 1 to 7:  0.617 0.661 0.687 0.689 0.689 0.688 0.687
```

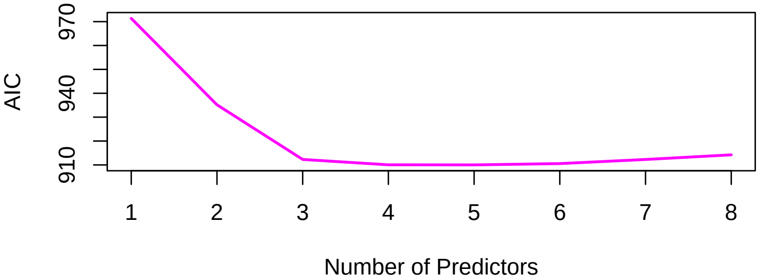


- Based on the **AIC vs p** plot, we see that for **4** regressors, the **AIC** is minimum.
- corresponding to **4**, we have **humidity**, **ibh**, **temp** and **vis** as regressors.

```
##
## Call:
## lm(formula = O3 ~ humidity + temp + ibh + vis, data = ozone[c(1:300),
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1978  -3.0437  -0.4037   2.7905  13.5956
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.3263210  1.9195018  -4.338 1.98e-05 ***
## humidity     0.0666379  0.0151839   4.389 1.59e-05 ***
## temp         0.3221667  0.0221297  14.558  < 2e-16 ***
## ibh         -0.0010325  0.0001766  -5.845 1.34e-08 ***
## vis         -0.0066438  0.0038770  -1.714   0.0876 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.529 on 295 degrees of freedom
## Multiple R-squared:  0.6934, Adjusted R-squared:  0.6892
## F-statistic: 166.8 on 4 and 295 DF,  p-value: < 2.2e-16
```

Model B

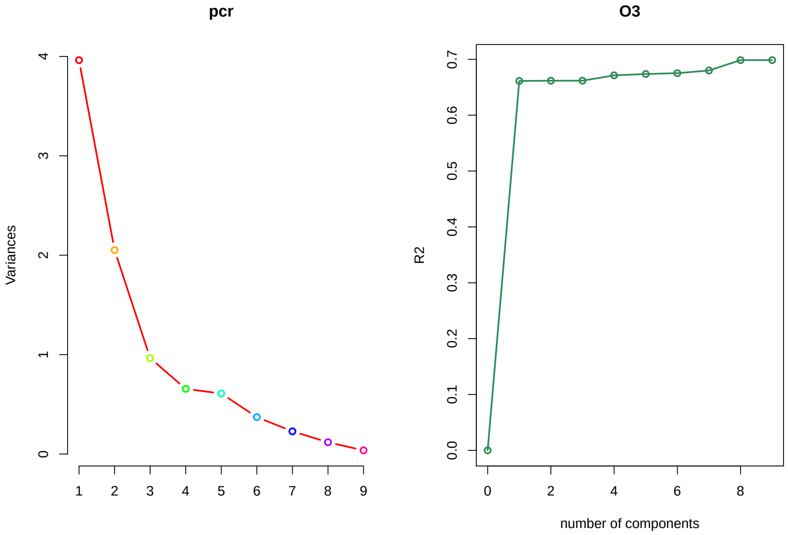
```
##   (Intercept)    vh  wind humidity temp   ibh   dpg   vis   doy   ibt
## 1        TRUE FALSE FALSE    FALSE TRUE FALSE FALSE FALSE FALSE FALSE
## 2        TRUE FALSE FALSE    FALSE TRUE  TRUE FALSE FALSE FALSE FALSE
## 3        TRUE FALSE FALSE     TRUE TRUE  TRUE FALSE FALSE FALSE FALSE
## 4        TRUE FALSE FALSE     TRUE TRUE  TRUE FALSE FALSE FALSE  TRUE
## 5        TRUE FALSE FALSE     TRUE TRUE  TRUE FALSE  TRUE FALSE  TRUE
## 6        TRUE  TRUE FALSE     TRUE TRUE  TRUE FALSE  TRUE FALSE  TRUE
## 7        TRUE  TRUE FALSE     TRUE TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## 8        TRUE  TRUE  TRUE     TRUE TRUE  TRUE FALSE  TRUE  TRUE  TRUE

## Mallows Cp value for p in 1 to 8:  71.593 29.682 5.912 3.723 3.741 4.333 6.06 8.007

## Adjusted R^2 value for p in 1 to 8:  0.617 0.661 0.687 0.691 0.692 0.692 0.691 0.69
```



Model B: Summary

```
##
## Call:
## lmridge.default(formula = O3 ~ humidity + temp + ibh + vis, data = ozone[1:300,
##     ], K = 0.018)
##
##
## Coefficients: for Ridge parameter K= 0.018
##             Estimate Estimate (Sc) StdErr (Sc) t-value (Sc) Pr(>|t|)
## Intercept    -7.8461    76763.2021  13502.8195       5.6850   <2e-16 ***
## humidity      0.0666       22.0888      4.9074       4.5011   <2e-16 ***
## temp          0.3154       80.5392      5.4528      14.7704   <2e-16 ***
## ibh          -0.0010      -32.0672      5.2791      -6.0743   <2e-16 ***
## vis          -0.0070       -9.4865      5.1123      -1.8556   0.0645 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Ridge Summary
##         R2     adj-R2   DF ridge           F         AIC         BIC
##    0.67850    0.67530    3.90232  167.29854  909.23688 2634.82497
## Ridge minimum MSE= 111.0194 at K= 0.018
## P-value for F-test ( 3.90232 , 296.0027 ) = 1.116608e-73
## -------------------------------------------------------------------
```

Model C: Scree Plot and Validation Plot



Model C: Remarks and R^2

- **scree-plot** gives us the indication of taking the first 4 PCs, as the elbow formation occurs at the 4^{th} PC till the 5^{th} PC.
- **validation plot**(validated by R^2) shows the cumulative amount of variation in Y explained by the PCs is mostly done by the first PC, with a slight increase with all the first 4 PCs.

```
## The value of R^2 taking first 4 PCs is :  0.6712925
```

Heteroscedasticity, Normality and Autocorrelation of Errors

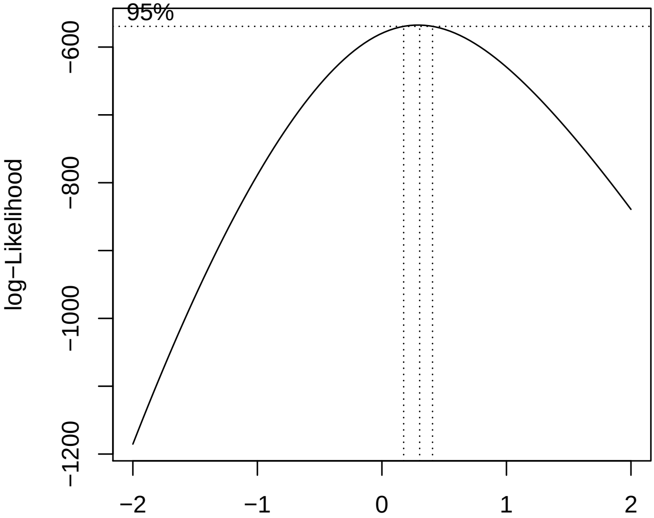
Heteroscedasticity of Errors: Breusch-Pagan(BP) Test and Box-Cox Transformation

Model A: BP Test

```
##
##  studentized Breusch-Pagan test
##
## data:  lmodA
## BP = 30.654, df = 4, p-value = 3.601e-06
```

- the test gets rejected i.e. the *errors are not homoscedastic* based on the data.

Model A: Box-Cox Transform



Model A: BP Test and Summary of transformed model

```
##
## Call:
## lm(formula = ((O3^lambdaA - 1)/lambdaA) ~ humidity + temp + ibh +
##     vis, data = ozone[1:300, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36633 -0.48378  0.04014  0.52043  2.12105
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.472e-01  3.242e-01  -0.454    0.650
## humidity     1.109e-02  2.564e-03   4.326 2.08e-05 ***
## temp         5.748e-02  3.737e-03  15.379  < 2e-16 ***
## ibh         -2.179e-04  2.983e-05  -7.305 2.60e-12 ***
## vis         -1.051e-03  6.548e-04  -1.606    0.109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7649 on 295 degrees of freedom
## Multiple R-squared:  0.7263, Adjusted R-squared:  0.7226
## F-statistic: 195.7 on 4 and 295 DF,  p-value: < 2.2e-16
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lmodA
## BP = 7.8305, df = 4, p-value = 0.09799
```

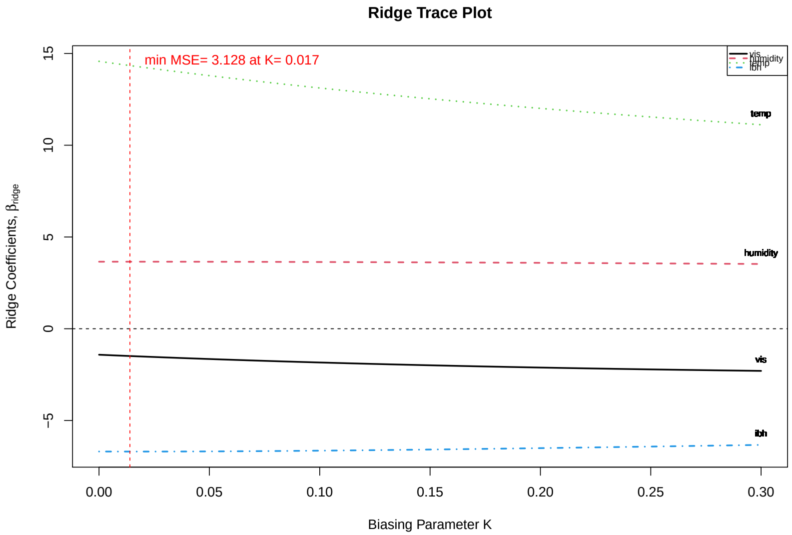
- The transformed model exhibits *homoscedasticity*

Model B: BP Test

```
##
##  studentized Breusch-Pagan test
##
## data:  lmodB
## BP = 30.654, df = 4, p-value = 3.601e-06
```

- the test gets rejected i.e. the *errors are not homoscedastic* based on the data.

Model B: Box-Cox Transform and Ridge complexity Parameter



Model B: BP Test and Summary of transformed model

```
##
## Call:
## lmridge.default(formula = ((O3^lambdaB - 1)/lambdaB) ~ vis +
##     humidity + temp + ibh, data = ozone[1:300, ], K = 0.017)
##
##
## Coefficients: for Ridge parameter K= 0.017
##             Estimate Estimate (Sc) StdErr (Sc) t-value (Sc) Pr(>|t|)
## Intercept    -0.0603    16176.0867   2268.7196       7.1301   <2e-16 ***
## vis          -0.0011       -1.5016      0.8588      -1.7485   0.0814 .
## humidity      0.0110        3.6527      0.8242       4.4317   <2e-16 ***
## temp          0.0560       14.2928      0.9163      15.5987   <2e-16 ***
## ibh          -0.0002       -6.6930      0.8870      -7.5457   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Ridge Summary
##         R2     adj-R2   DF ridge          F        AIC        BIC
##    0.71170    0.70870    3.90756  196.23452 -162.00036 1563.60714
## Ridge minimum MSE= 3.128213 at K= 0.017
## P-value for F-test ( 3.90756 , 296.0025 ) = 6.065563e-81
## -------------------------------------------------------------------
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lmodB
## BP = 7.9005, df = 4, p-value = 0.09529
```

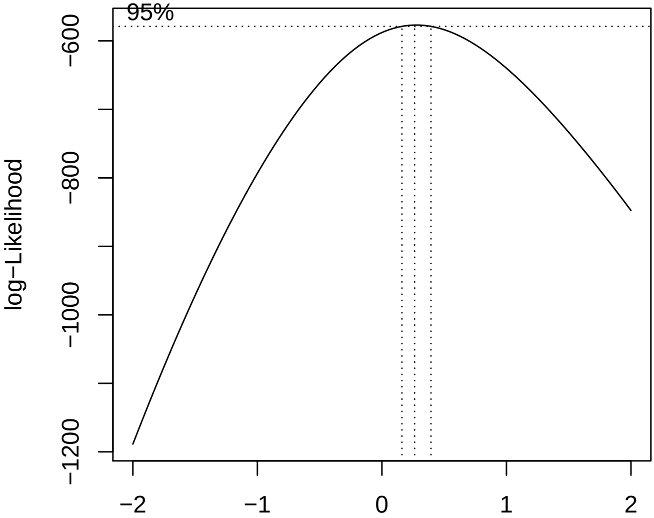
- The transformed model exhibits *homoscedasticity*

Model C: BP Test

```
##
##  studentized Breusch-Pagan test
##
## data:  lmodC
## BP = 30.719, df = 4, p-value = 3.494e-06
```

- the test gets rejected i.e. the *errors are not homoscedastic* based on the data.

Model C: Box-Cox Transform



Model C: BP Test and R^2 of transformed model

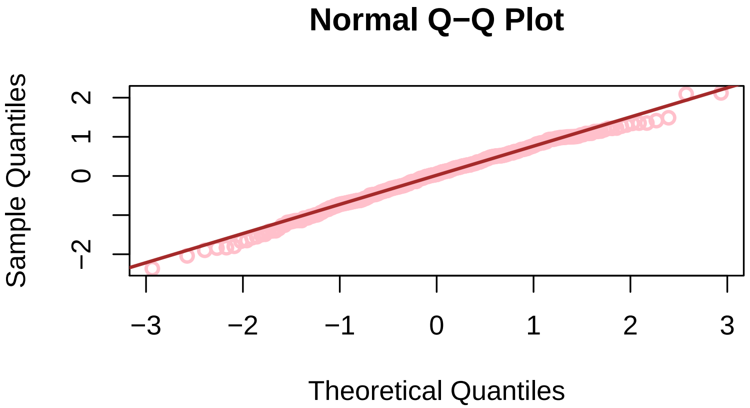
```
##
##  studentized Breusch-Pagan test
##
## data:  lmodA
## BP = 7.8305, df = 4, p-value = 0.09799
```

```
## The R^2 value of the transformed model is :  0.7252028
```

- The transformed model exhibits *homoscedasticity*

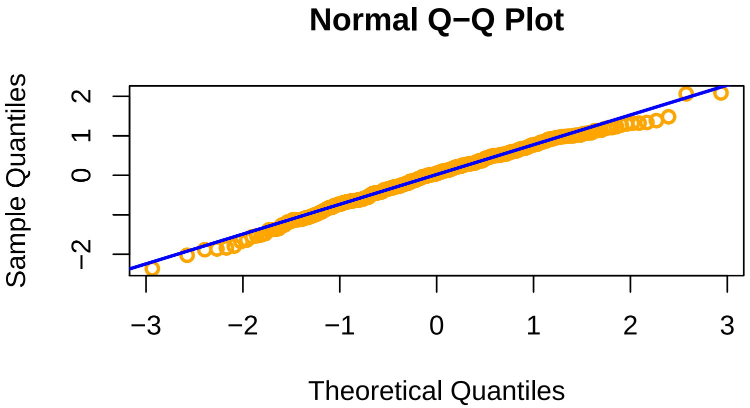
Normality of Errors

Model A: Normal Q-Q Plot and Shapiro-Wilks Test



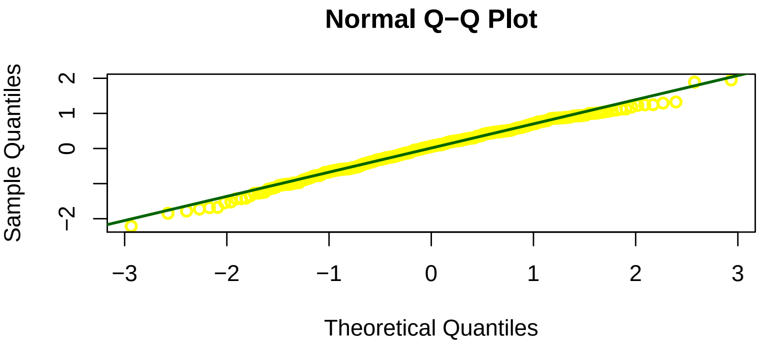

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lmodA)
## W = 0.99233, p-value = 0.1246
```

Model B: Normal Q-Q Plot and Shapiro-Wilks Test




```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lmodB)
## W = 0.99191, p-value = 0.1007
```

Model C: Normal Q-Q Plot and Shapiro-Wilks Test



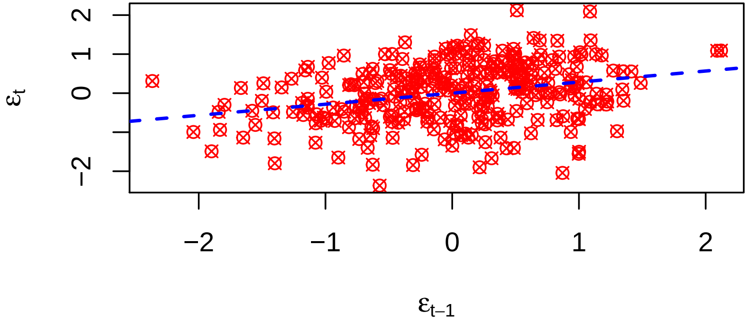

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lmodC)
## W = 0.99121, p-value = 0.07045
```

- The errors are normally distributed based on the data and the above models

Autocorrelation of Errors

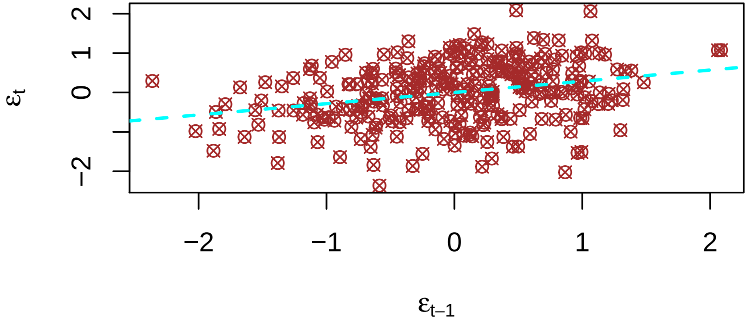
Detection of Autocorrelation: ϵ_t *vs.* ϵ_{t-1} Plot and Durbin-watson(DW) Test

Model A



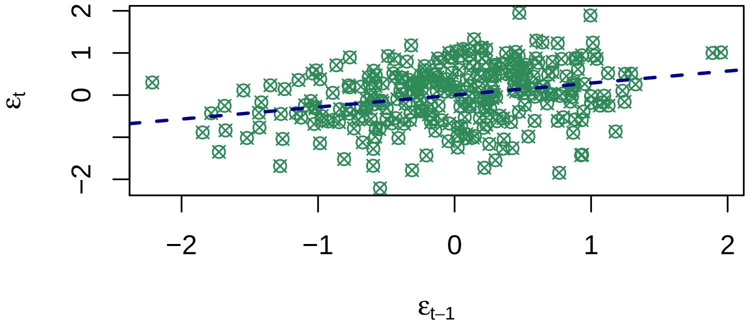

```
##
##  Durbin-Watson test
##
## data:  lmodA
## DW = 1.4316, p-value = 2.075e-07
## alternative hypothesis: true autocorrelation is greater than 0
```

Model B




```
##
##  Durbin-Watson test
##
## data:  lmodB
## DW = 1.4314, p-value = 2.054e-07
## alternative hypothesis: true autocorrelation is greater than 0
```

Model C

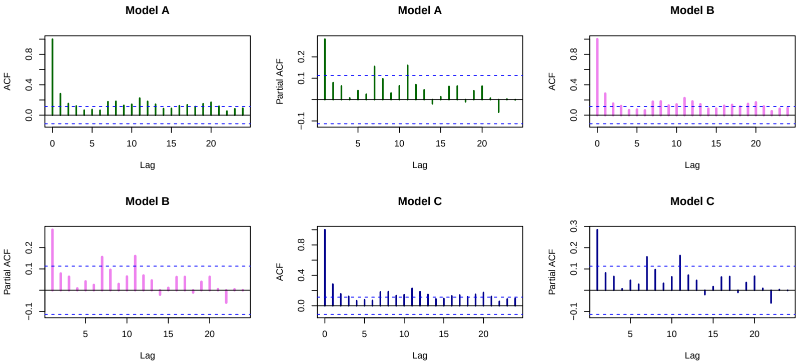



```
##
##  Durbin-Watson test
##
## data:  lmodC
## DW = 1.4288, p-value = 1.824e-07
## alternative hypothesis: true autocorrelation is greater than 0
```

Correction for Autocorrelation

AR(p) Errors and ACF and PACF Plots

- Assuming **AR(p)** model for the errors, we fitted models for *p=1-20*. None performed satisfactorily i.e. none achieved stationarity.
- We look at the **acf** and the **pacf** plots of the residuals of each model to see if $AR(p)$ is indeed a good error model
- $AR(p)$ model does not seem to be a good model for the errors.



Model A

- we use the **auto.arima** function in the **forecast** package in **R** that automatically fits an **ARIMA(p,d,q)** process by taking that value of **d** such that **stationarity is achieved** and **p** and **q** are chosen so that minimum **AIC** is achieved.

```
## Series: (ozone[c(1:300), 1]^lambdaA - 1)/lambdaA
## Regression with ARIMA(0,1,1) errors
##
## Coefficients:
##           ma1   drift  humidity    temp      ibh      vis
##       -0.9155  0.0018    0.0050  0.0581  -2e-04  -0.0019
## s.e.   0.0244  0.0025    0.0027  0.0045   0e+00   0.0006
##
## sigma^2 estimated as 0.5212:  log likelihood=-324.73
## AIC=663.47   AICc=663.85   BIC=689.37

## The R^2 value of modA is :  0.7662688
```

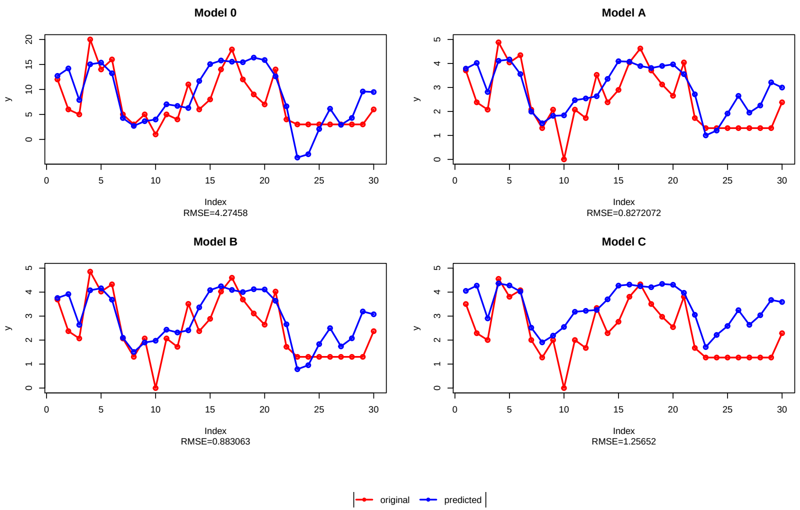
Remarks for Model A, B and C

- In model **A**, an **ARIMA(0,1,2)** model is fitted.
- We do not take any remedial measure for model **B** and **C** as the problem then becomes too complicated.
- Possibly better models may be fitted after a course on *Time Series Analysis*.

Prediction

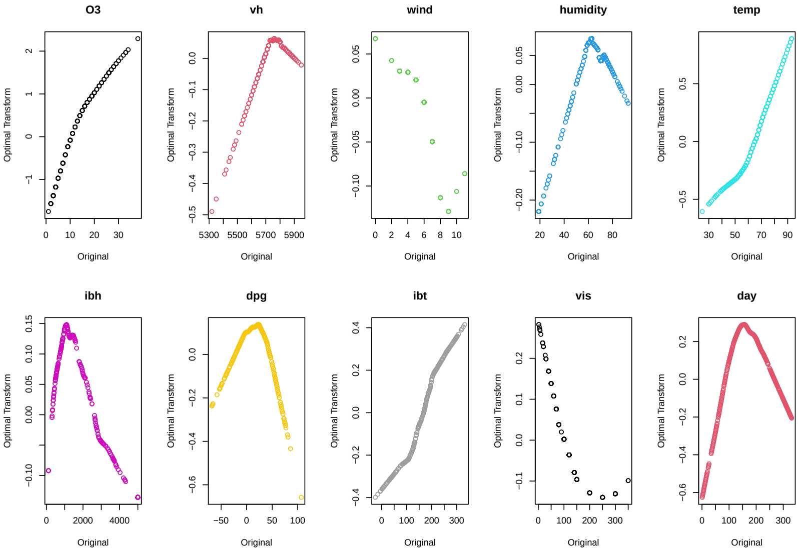
Prediction

- based on the **RMSE values**, model **A** performs best
- model **B** is a close competitor.
- Model **C** performs comparatively poor - a model without autocorrelation correction may be a reason.



Alternating Conditonal Expectation

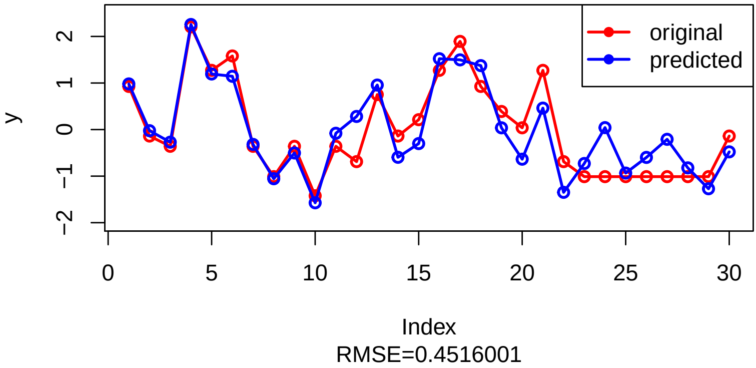
Optimal Transformations



ACE Model and Summary

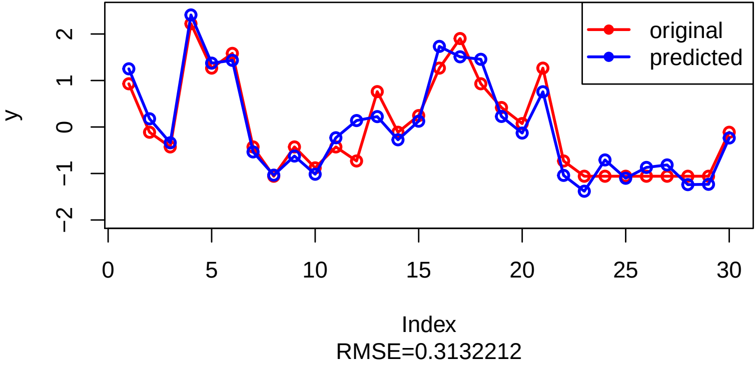
```
##
## Call:
## lm(formula = O3 ~ ., data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26357 -0.23023  0.02591  0.29252  0.99866
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.313e-16  2.344e-02   0.000 1.000000
## vh           1.220e+00  3.260e-01   3.744 0.000218 ***
## wind         1.693e+00  5.495e-01   3.081 0.002262 **
## humidity     7.432e-01  3.152e-01   2.358 0.019050 *
## temp         8.979e-01  1.412e-01   6.361 7.77e-10 ***
## ibh          7.366e-01  3.471e-01   2.122 0.034657 *
## dpg          1.388e+00  1.858e-01   7.468 9.61e-13 ***
## ibt          1.031e+00  2.772e-01   3.720 0.000239 ***
## vis          1.285e+00  2.412e-01   5.328 1.99e-07 ***
## day          1.347e+00  1.273e-01  10.581  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4061 on 290 degrees of freedom
## Multiple R-squared:  0.8406, Adjusted R-squared:  0.8357
## F-statistic: 169.9 on 9 and 290 DF,  p-value: < 2.2e-16
```

Prediction based on ACE Model



Final ACE Model

- we have seen that **ibt** and **temp** are almost perfectly correlated and **vh** showed a similar relationship with either of them.
- We again fit a linear model, **Ace**, based on the transformed data, removing **ibt** and **vh**.




```
## The R-squared value of the final model is:  0.8271309
```

Conclusion

- with **Model 0** as baseline, the R^2 value and the **RMSE** value of **Model 0**, **Model A**, **Model B**, **Model C** and **ACE** model are compared.

Model type	Model Name	R^2	RMSE
Parametric	Model 0	0.6986	4.2745
	Model A	0.7662	0.8272
	Model B	0.7202	0.8830
	Model C	0.7077	1.2565
Non-Parametric	ACE	0.8271	0.3132

- Among the **parametric models**, **model A** has the **highest** R^2 value as well as the **lowest** *RMSE* value.
- All models - **A**, **B** and **C** are better than the baseline model **Model 0**. This validates our corrections for **multicollinearity**, **heteroscedasticity** and **autocorrelation** and **variable selection**.
- Simple **non-parametric models** are better if the problem of prediction is to be solved. But here, the **ACE** model transforms the data so that maximum R^2 can be achieved. And, as expected it has the **highest** R^2 value and the **lowest** *RMSE* value amond all the models.
- So among the models considered here, **ACE** model is the **best**, both for the problem of prediction and for the purpose of explaining **ozone concentration** by the **meteorological** variables based on the **ozone** dataset.
- The entire project along with source code is available at : *https://github.com/ArkaB-DS/Modelling-linear-relationship-between-Ozone-Concentration-and-Meteorology-LA-Basin-1976*

Bibliography

1. Leo Breiman & Jerome H. Friedman (1985): Estimating Optimal Transformations for Multiple Regression and Correlation, Journal of the American Statistical Association, 80:391, 580-598
2. Jolliffe, Ian T. (1982). "A note on the Use of Principal Components in Regression". Journal of the Royal Statistical Society, Series C. 31 (3): 300–303. doi:10.2307/2348005. JSTOR 2348005.
3. Sung H. Park (1981). "Collinearity and Optimal Restrictions on Regression Parameters for Estimating Responses". Technometrics. 23 (3): 289–295. doi:10.2307/1267793.
4. Wilkinson, L., & Dallal, G.E. (1981). Tests of significance in forward selection regression with an F-to enter stopping rule. Technometrics, 23, 377–380
5. Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle", in Petrov, B. N.; Csáki, F. (eds.), 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadémiai Kiadó, pp. 267–281. Republished in Kotz, S.; Johnson, N. L., eds. (1992), Breakthroughs in Statistics, I, Springer-Verlag, pp. 610–624.
6. Akaike, H. (1974), "A new look at the statistical model identification", IEEE Transactions on Automatic Control, 19 (6): 716–723, doi:10.1109/TAC.1974.1100705, MR 0423716.
7. Shapiro, S. S.; Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)". Biometrika. 52 (3–4): 591–611. doi:10.1093/biomet/52.3-4.591. JSTOR 2333709. MR 0205384. p. 593
8. Breusch, T. S.; Pagan, A. R. (1979). "A Simple Test for Heteroskedasticity and Random Coefficient Variation". Econometrica. 47 (5): 1287–1294. doi:10.2307/1911963. JSTOR 1911963. MR 0545960.
9. Box, George E. P.; Cox, D. R. (1964). "An analysis of transformations". Journal of the Royal Statistical Society, Series B. 26 (2): 211–252. JSTOR 2984418. MR 0192611.
10. Durbin, J.; Watson, G. S. (1950). "Testing for Serial Correlation in Least Squares Regression, I". Biometrika. 37 (3–4): 409–428. doi:10.1093/biomet/37.3-4.409. JSTOR 2332391
11. Durbin, J.; Watson, G. S. (1951). "Testing for Serial Correlation in Least Squares Regression, II". Biometrika. 38 (1–2): 159–179. doi:10.1093/biomet/38.1-2.159. JSTOR 2332325
12. Faraway, J.J. (2004). Linear Models with R (1st ed.). Chapman and Hall/CRC. https://doi.org/10.4324/9780203507278
13. Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. (1975). Ridge regression: Some simulations. Communications in Statistics-Theory and Methods, 4(2), 105-123.