CARINNA ROEBUCK &
SIDNEY CHRISTENSEN

LA OZONE

Determining Atmospheric Concentration

BACKGROUND

Ozone pollution is a byproduct of coal and fossil fuel burning, occurring when internal combustion engines and fossil fuel power plants release volatile organic compounds and nitrogen oxide that interact with oxygen and UV rays.

The aim of this study is to examine eight of the most suspected ozone concentration indicators/predictors and evaluate the significance of each. The concentrations in the data are sampled from Upland, California with each sample representing one day of the 330 days sampled throughout the year of 1976.

DATA OVERVIEW

Data Source

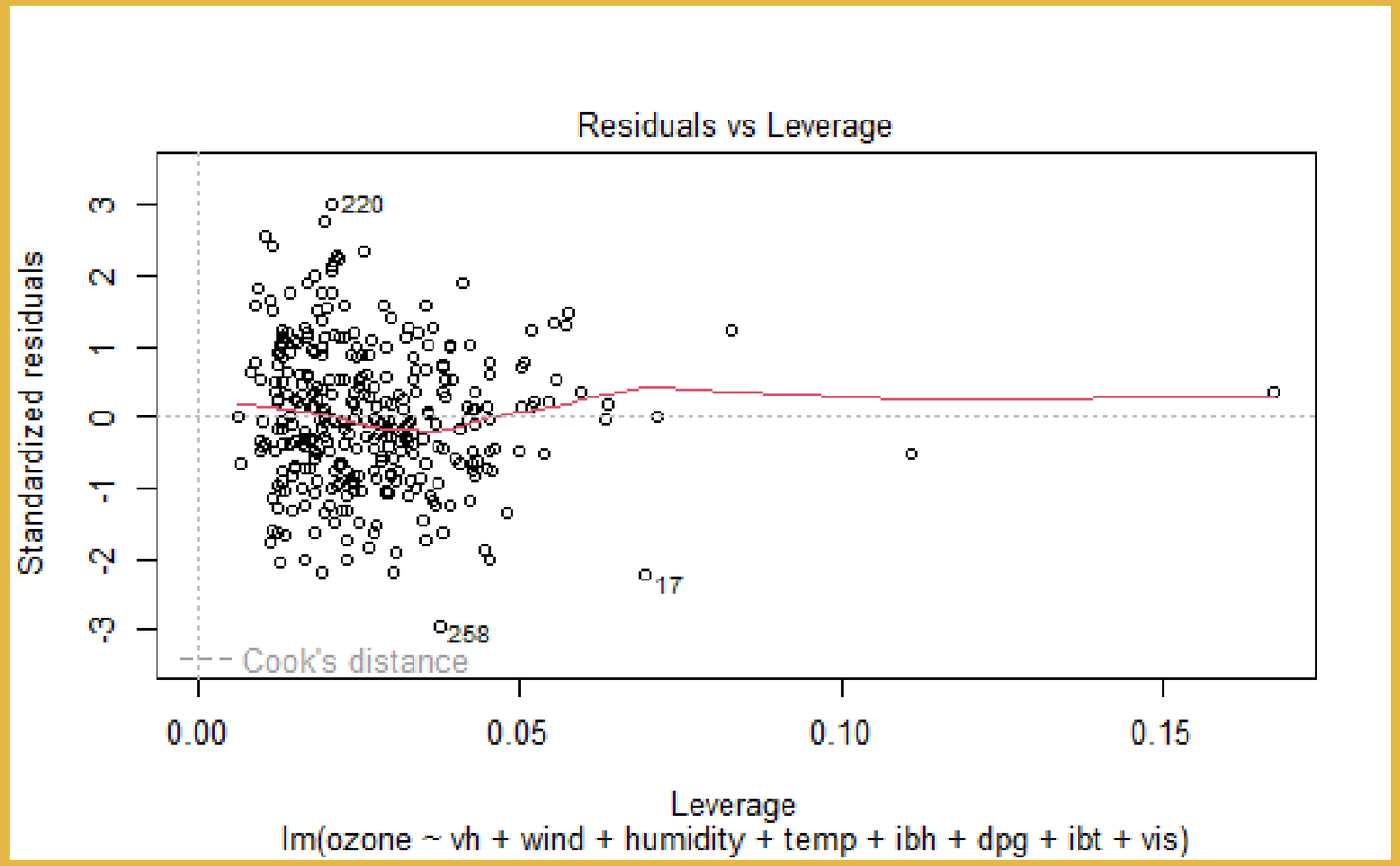
Elements of Stats Learning by Trevor Hastie with study conducted by Leo Breiman

330
Observations

Target variable

Upland Maximum Daily Ozone

& 8 Numerical Predictors



Missing Data

There were no 0s or missing data in this dataset

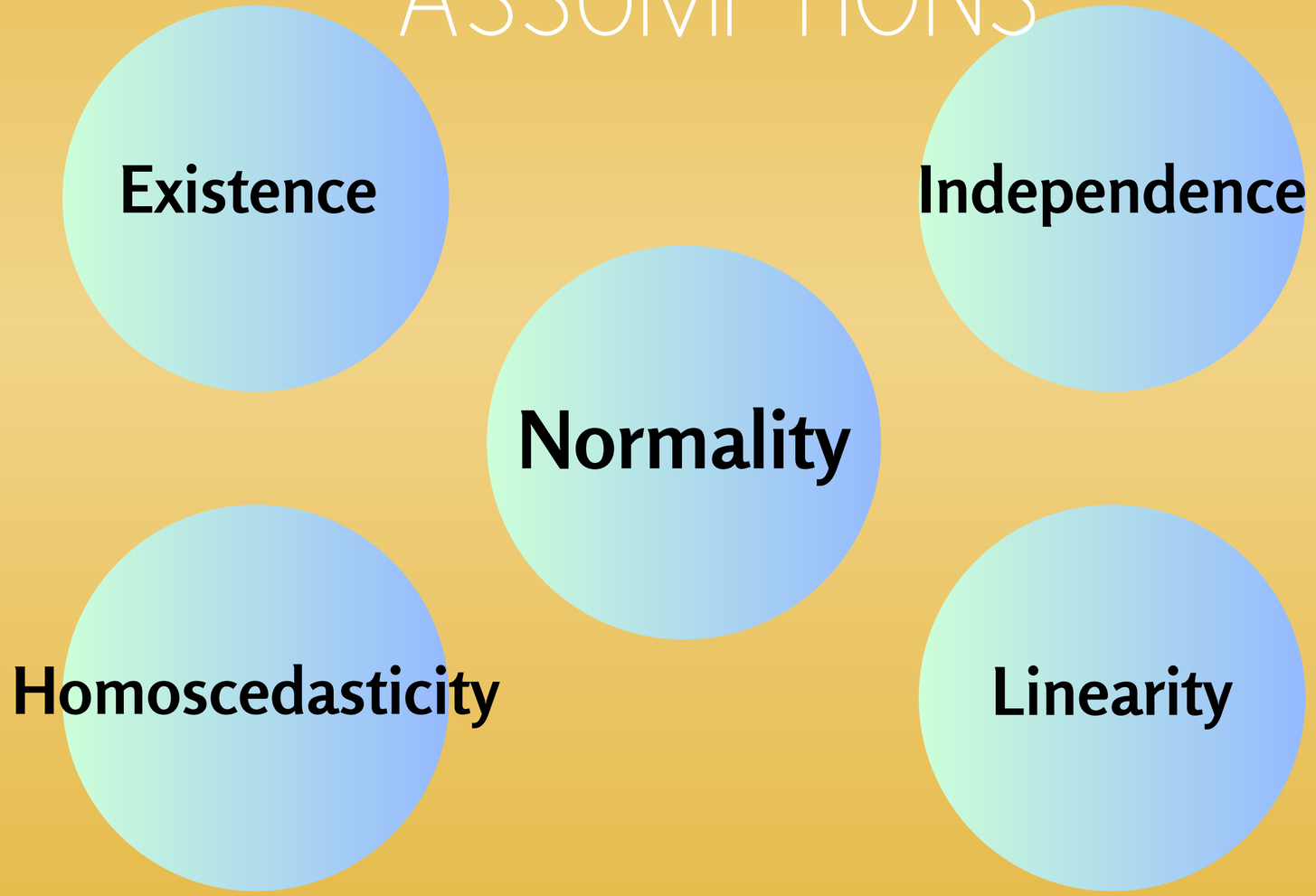
DATA PREPARATION

PREDICTORS

A psych: 8 × 13

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
vh	1	330	5750.484848	105.708241	5760.0	5759.848485	103.7820	5320	5950	630	-0.956917847	1.38327153	5.8190472
wind	2	330	4.890909	2.293159	5.0	4.829545	1.4826	0	21	21	1.062130828	6.83616972	0.1262343
humidity	3	330	58.130303	19.865000	64.0	59.981061	14.8260	19	93	74	-0.812845941	-0.48189560	1.0935323
temp	4	330	61.754545	14.458737	62.0	61.780303	16.3086	25	93	68	0.008176912	-0.61717329	0.7959273
ibh	5	330	2572.875758	1803.885870	2112.5	2546.140152	2225.3826	111	5000	4889	0.273816685	-1.54804390	99.3006489
dpg	6	330	17.369697	35.717181	24.0	18.746212	38.5476	-69	107	176	-0.302083583	-0.64912041	1.9661661
ibt	7	330	161.160606	76.679424	167.5	162.681818	78.5778	-25	332	357	-0.168381600	-0.61377630	4.2210633
vis	8	330	124.533333	79.362393	120.0	116.174242	74.1300	0	350	350	0.822248287	0.04113327	4.3687560

LINEAR REGRESSION ASSUMPTIONS



EXISTENCE

For any fixed value of the variable X, there exists a random variable Y with a certain probability distribution having finite mean $(\mu Y|X)$ and variance $(\sigma Y|X2$).

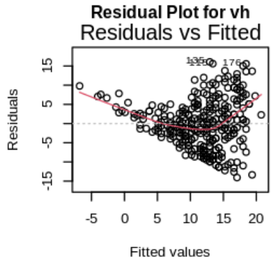
All regression models satisfy this assumption regardless if they are linear or not

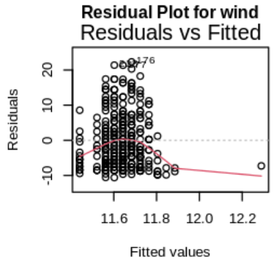
	vars	n
vh	1	330
wind	2	330
humidity	3	330
temp	4	330
ibh	5	330
dpg	6	330
ibt	7	330
vis	8	330

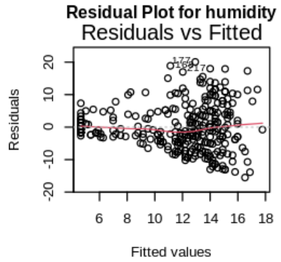
All observations are only made once during the 330 days of the study and only one recorded value for each predictor, so we can assume the Y values are independent.

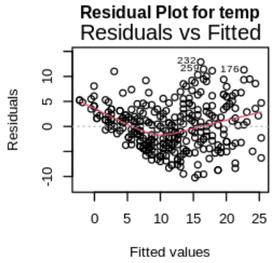
INDEPENDENCE

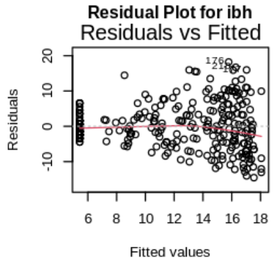
LINEARITY

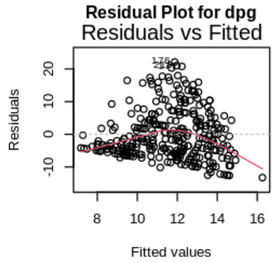


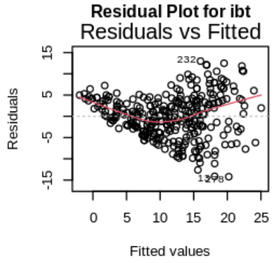


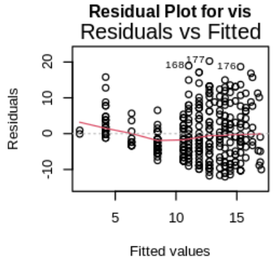




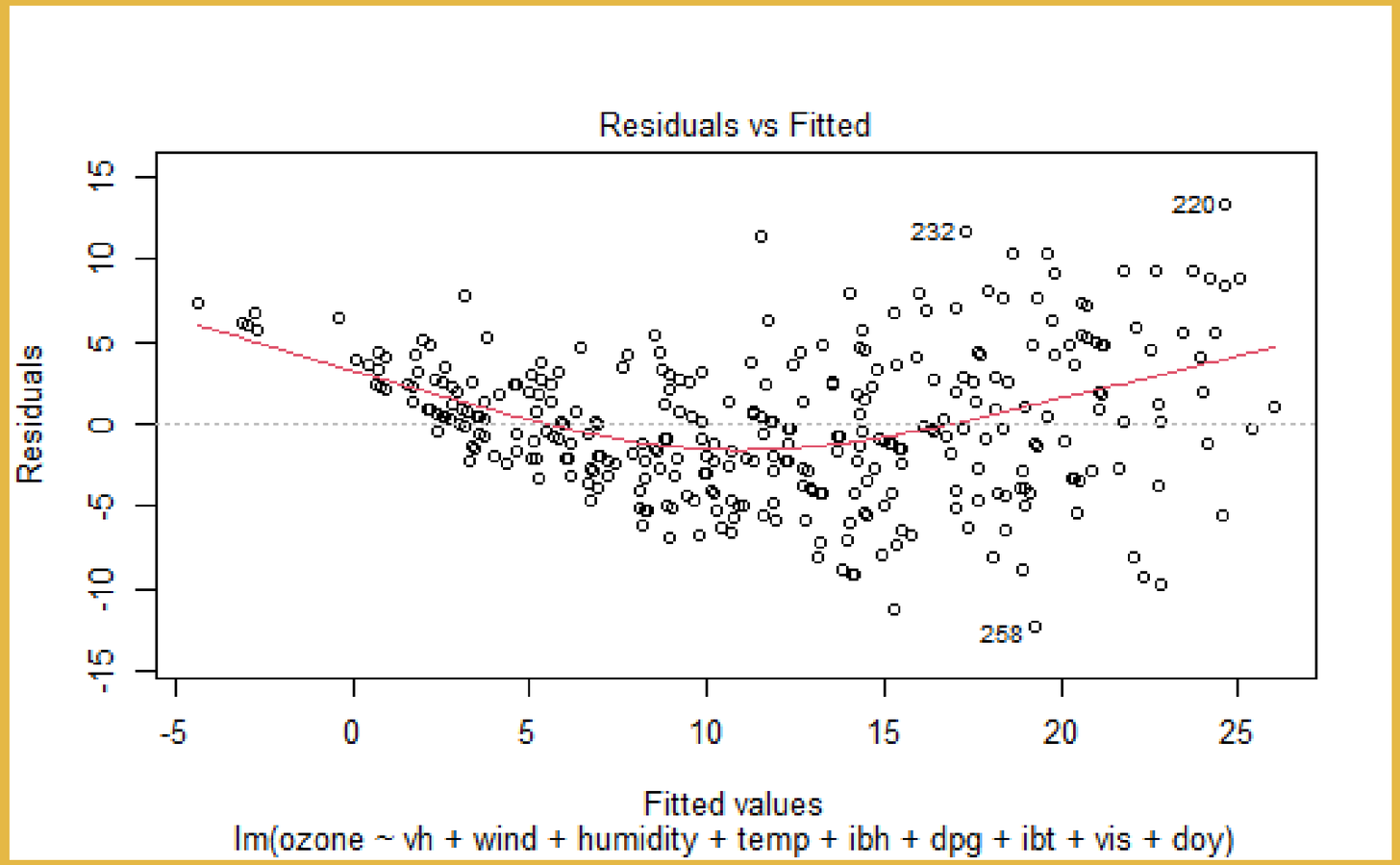






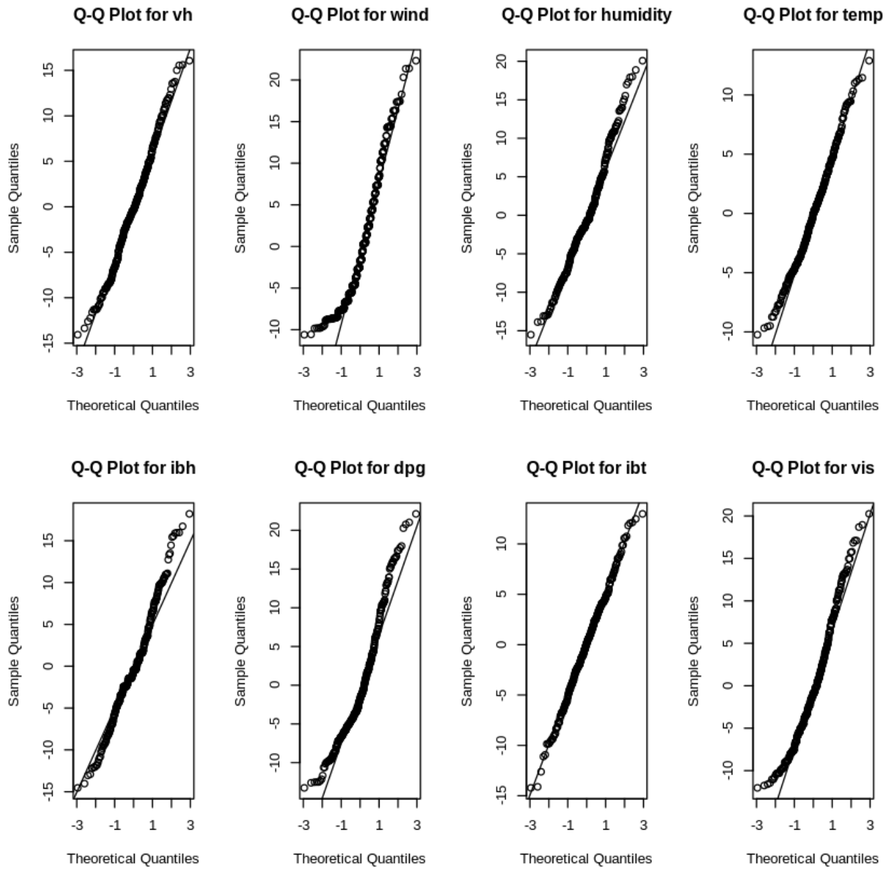


All of the predictor variables show some level of curvilinear nature, violating the linearity assumption for linear regression.



HOMOSCEDASTICITY

NORMALITY



Utilizing the QQ-Plot, we observe normality with some problem variables evident

INITIAL MODEL:

Upland Maximum Ozone = β0+β1{vh}+β2{wind}+β3{humidity}+β4{temp}+β5{ibh}+β6{dpg}+β7{ibt}+β8{vis}+ε

```
Call:
lm(formula = target ~ vh + temp + wind + ibh + ibt + vis + humidity +
    dpg, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5545 -2.8701 -0.1397  2.5864 11.3276

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.6529192 28.2145325   1.086   0.2781
vh          -0.0076034  0.0051531  -1.476   0.1411
temp         0.3325338  0.0490773   6.776 6.39e-11 ***
wind         0.0438064  0.1138249   0.385   0.7006
ibh         -0.0006923  0.0002675  -2.588   0.0101 *
ibt          0.0122363  0.0133687   0.915   0.3608
vis         -0.0061194  0.0034530  -1.772   0.0774 .
humidity     0.0827900  0.0176843   4.682 4.29e-06 ***
dpg         -0.0120860  0.0108968  -1.109   0.2682
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.06 on 305 degrees of freedom
Multiple R-squared:  0.7304,    Adjusted R-squared:  0.7233
F-statistic: 103.3 on 8 and 305 DF,  p-value: < 2.2e-16
R-squared: 0.7303855
Mean Squared Error (MSE): 16.01253
```

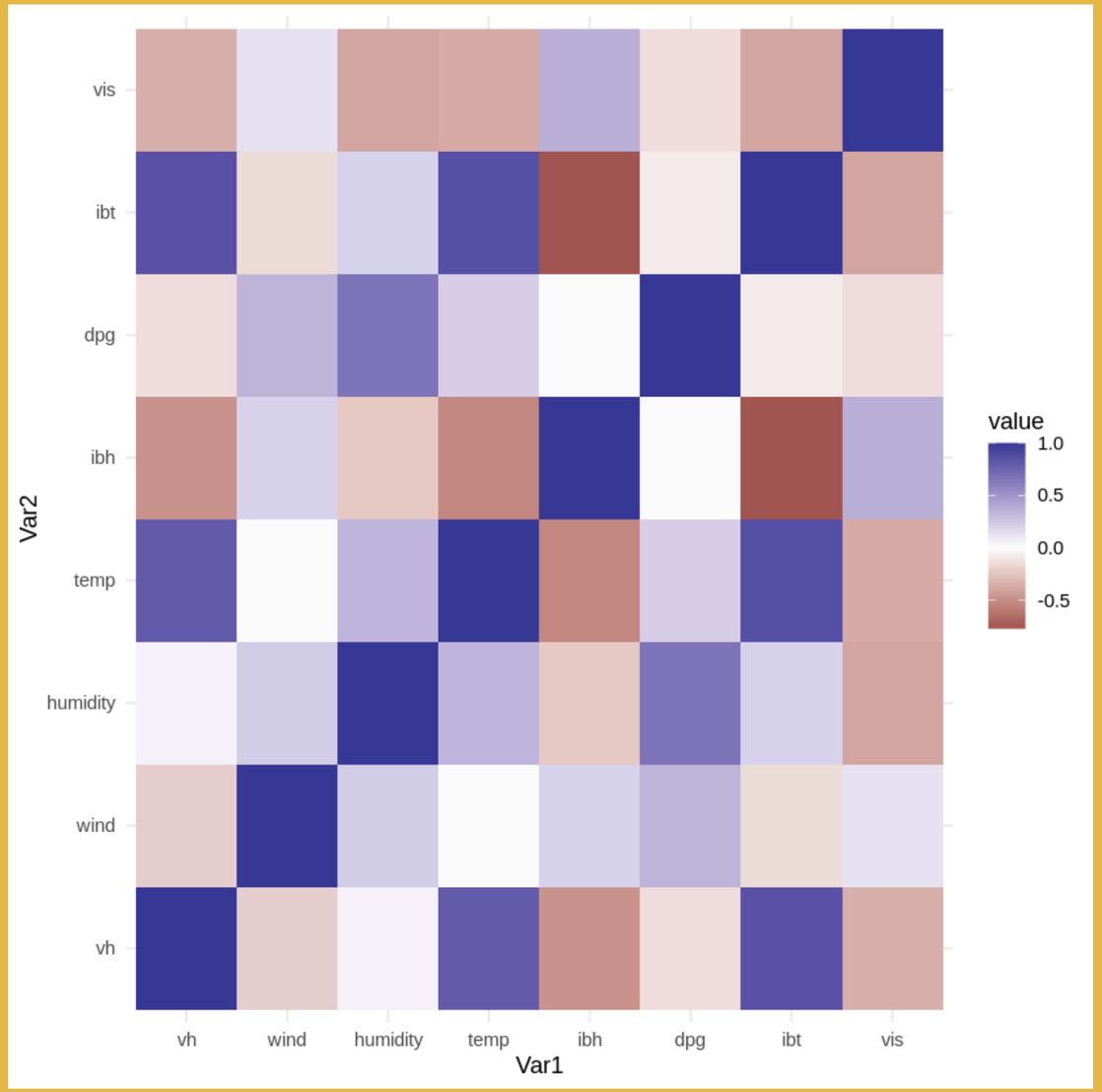
CORRELATION & COLLINEARITY

	vh	wind	humidity	temp	ibh	dpg	ibt	vis
vh	1.00000000	-0.203534497	0.05297558	0.799281363	-0.489466795	-0.141501517	0.84975890	-0.3411195
wind	-0.20353450	1.000000000	0.23504633	0.008837492	0.194253184	0.347146539	-0.15267535	0.1336195
humidity	0.05297558	0.235046331	1.00000000	0.337636511	-0.239880249	0.679434787	0.18935029	-0.3835856
temp	0.79928136	0.008837492	0.33763651	1.000000000	-0.530665582	0.214638326	0.86853462	-0.3746866
ibh	-0.48946680	0.194253184	-0.23988025	-0.530665582	1.000000000	0.006313991	-0.76937856	0.3726626
dpg	-0.14150152	0.347146539	0.67943479	0.214638326	0.006313991	1.000000000	-0.07190331	-0.1447323
ibt	0.84975890	-0.152675347	0.18935029	0.868534617	-0.769378556	-0.071903310	1.00000000	-0.4000139
vis	-0.34111951	0.133619492	-0.38358563	-0.374686647	0.372662568	-0.144732337	-0.40001386	1.0000000

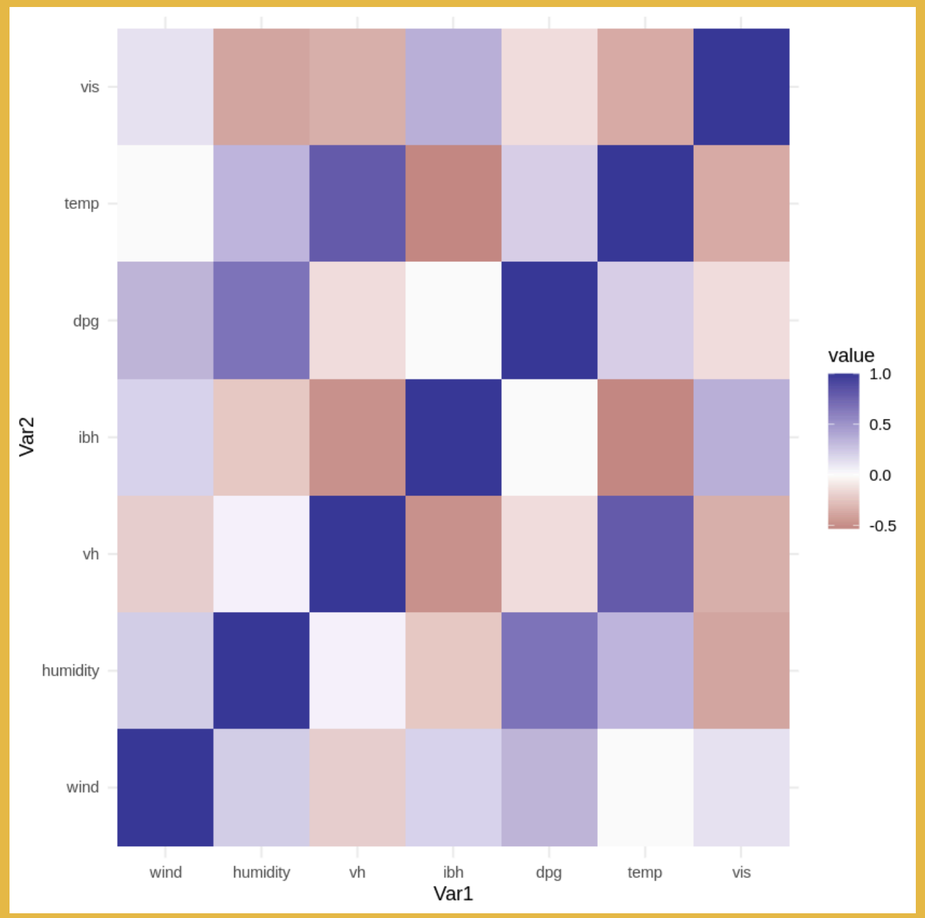
vh	temp	wind	humidity	ibh	ibt	dpg	vis
5.232404	9.202481	1.276787	2.369847	4.374409	19.031512	2.879601	1.415681

CORRELATION & COLLINEARITY

Here we can visualize the relationship between the variables. Evidently ibt and temp pose issues



CORRELATION & COLLINEARITY



	wind	humidity	vh	ibh	dpg	temp	vis
wind	1.000000000	0.23504633	-0.20353450	0.194253184	0.347146539	0.008837492	0.1336195
humidity	0.235046331	1.00000000	0.05297558	-0.239880249	0.679434787	0.337636511	-0.3835856
vh	-0.203534497	0.05297558	1.00000000	-0.489466795	-0.141501517	0.799281363	-0.3411195
ibh	0.194253184	-0.23988025	-0.48946680	1.000000000	0.006313991	-0.530665582	0.3726626
dpg	0.347146539	0.67943479	-0.14150152	0.006313991	1.000000000	0.214638326	-0.1447323
temp	0.008837492	0.33763651	0.79928136	-0.530665582	0.214638326	1.000000000	-0.3746866
vis	0.133619492	-0.38358563	-0.34111951	0.372662568	-0.144732337	-0.374686647	1.0000000

Notice how the correlation matrix becomes lighter in color after removing ibt

MODEL REDUCTION METHODS

Backwards Selection

Backwards selection is a stepwise regression approach that starts with a model including all independent variables and progressively removes the least significant ones

Stepwise Selection

A modified version of forward selection that allows for re-evaluation of variables after they've been added. All variables are checked again for effectiveness through a partial f-test

Lasso

LASSO regression shrinks down all coefficients towards zero, and if a predictor hits a certain threshold that coefficient will be set to zero and thus removed from the model.

Example: Stepwise

```
Start:  AIC=887.7
target ~ temp + wind + humidity + vh + ibh + dpg + vis

           Df Sum of Sq    RSS    AIC
- wind      1      3.41 5045.2 885.91
- vh        1     23.21 5065.0 887.14
<none>                  5041.7 887.70
- dpg       1     44.83 5086.6 888.48
- vis       1     53.14 5094.9 888.99
- humidity  1    371.27 5413.0 908.01
- ibh       1    495.72 5537.5 915.15
- temp      1   1861.79 6903.5 984.38
```

```
Step:  AIC=885.12
target ~ temp + humidity + ibh + vis

           Df Sum of Sq    RSS     AIC
<none>                  5097.0  885.12
- vis       1      49.9 5146.9  886.18
- humidity  1     529.7 5626.7  914.17
- ibh       1     570.2 5667.3  916.42
- temp      1    4480.3 9577.3 1081.18
```

BACKWARDS SELECTION

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	wind	0.7295	0.7242	6.2067	1779.0054	4.0539
2	vh	0.728	0.7236	5.8559	1778.6920	4.0582
3	dpg	0.7267	0.7231	5.3536	1778.2160	4.0614
4	vis	0.724	0.7213	6.3818	1779.2747	4.0747

Backward Elimination Method

Candidate Terms:

1 . temp
2 . wind
3 . humidity
4 . vh
5 . ibh
6 . dpg
7 . vis

We are eliminating variables based on p value...

Final Model Output

Model Summary

R	0.851	RMSE	4.075
R-Squared	0.724	Coef. Var	35.015
Adj. R-Squared	0.721	MSE	16.603
Pred R-Squared	0.717	MAE	3.230

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	13501.717	3	4500.572	271.072	0.0000
Residual	5146.895	310	16.603		
Total	18648.611	313			

Parameter Estimates

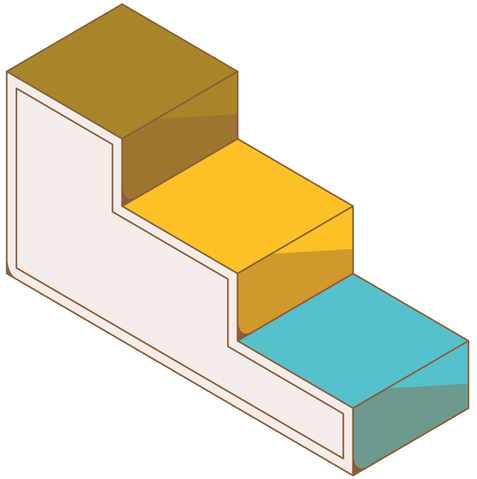
model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-11.089	1.511		-7.337	0.000	-14.063	-8.115
temp	0.335	0.020	0.615	16.899	0.000	0.296	0.374
humidity	0.079	0.012	0.204	6.410	0.000	0.055	0.103
ibh	-0.001	0.000	-0.225	-6.362	0.000	-0.001	-0.001

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.109e+01	1.511e+00	-7.337	1.93e-12	***
temp	3.348e-01	1.981e-02	16.899	< 2e-16	***
ibh	-9.663e-04	1.519e-04	-6.362	7.13e-10	***
humidity	7.874e-02	1.228e-02	6.410	5.40e-10	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

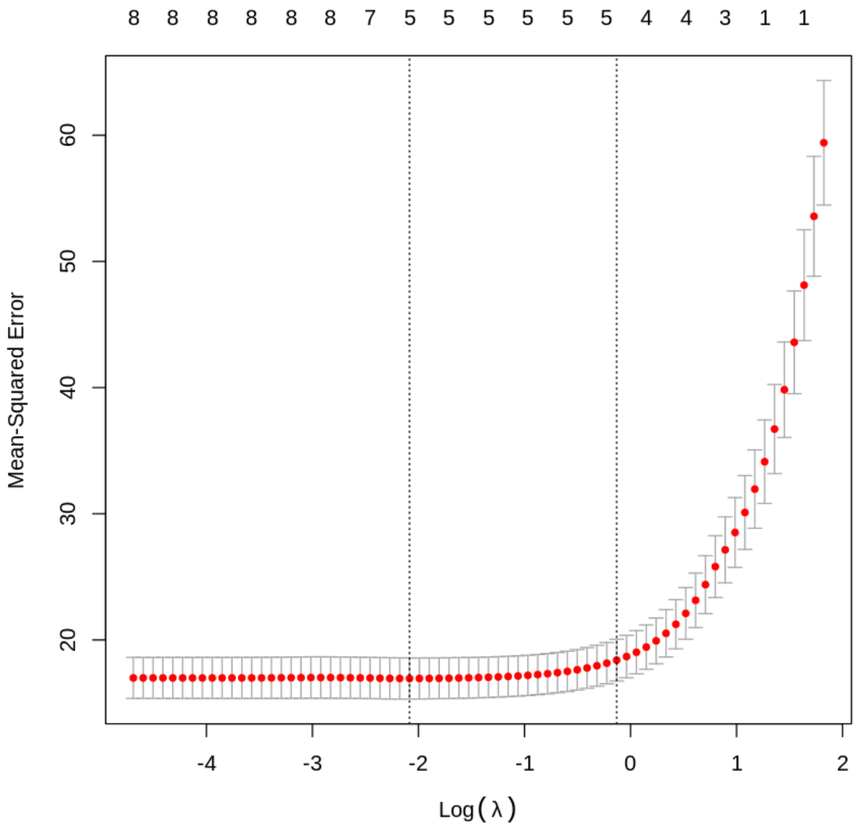
```
Start:  AIC=887.7
target ~ temp + wind + humidity + vh + ibh + dpg + vis

           Df Sum of Sq    RSS    AIC
- wind      1      3.41 5045.2 885.91
- vh        1     23.21 5065.0 887.14
<none>                  5041.7 887.70
- dpg       1     44.83 5086.6 888.48
- vis       1     53.14 5094.9 888.99
- humidity  1    371.27 5413.0 908.01
- ibh       1    495.72 5537.5 915.15
- temp      1   1861.79 6903.5 984.38
```



```
Step:  AIC=885.12
target ~ temp + humidity + ibh + vis

           Df Sum of Sq    RSS     AIC
<none>                  5097.0  885.12
- vis       1      49.9 5146.9  886.18
- humidity  1     529.7 5626.7  914.17
- ibh       1     570.2 5667.3  916.42
- temp      1    4480.3 9577.3 1081.18

Call:
lm(formula = target ~ temp + humidity + ibh + vis, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9254 -2.8163 -0.2613  2.7248 11.6006

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.8015285  1.6786745  -5.839 1.33e-08 ***
temp         0.3294559  0.0199904  16.481  < 2e-16 ***
humidity     0.0723948  0.0127757   5.667 3.34e-08 ***
ibh         -0.0009101  0.0001548  -5.880 1.07e-08 ***
vis         -0.0058266  0.0033502  -1.739    0.083 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.061 on 309 degrees of freedom
Multiple R-squared:  0.7267,    Adjusted R-squared:  0.7231
F-statistic: 205.4 on 4 and 309 DF,  p-value: < 2.2e-16
R-squared: 0.726682
Mean Squared Error (MSE): 16.23249
```

STEPWISE SELECTION

LASSO REGRESSION




```
Call:
lm(formula = target ~ temp + ibh + ibt + vis + humidity, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5816 -2.8515 -0.3052  2.7211 11.4688

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.003e+01  1.701e+00  -5.898 9.66e-09 ***
temp         2.986e-01  4.120e-02   7.248 3.44e-12 ***
ibh         -7.544e-04  2.388e-04  -3.159  0.00174 **
ibt          8.646e-03  1.009e-02   0.857  0.39236
vis         -5.437e-03  3.382e-03  -1.608  0.10896
humidity     7.760e-02  1.415e-02   5.483 8.72e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.063 on 308 degrees of freedom
Multiple R-squared:  0.7273,    Adjusted R-squared:  0.7229
F-statistic: 164.3 on 5 and 308 DF,  p-value: < 2.2e-16
R-squared: 0.7273315
Mean Squared Error (MSE): 16.19391
```

60.2% Backwards Selection

Stepwise Selection 68.7%

72.7% Lasso Regression

Ridge Regression 72.5%

MODEL SELECTION: DETERMINING THE BEST MODEL

MEAN SQUARED ERRORS

Initial: 16.012

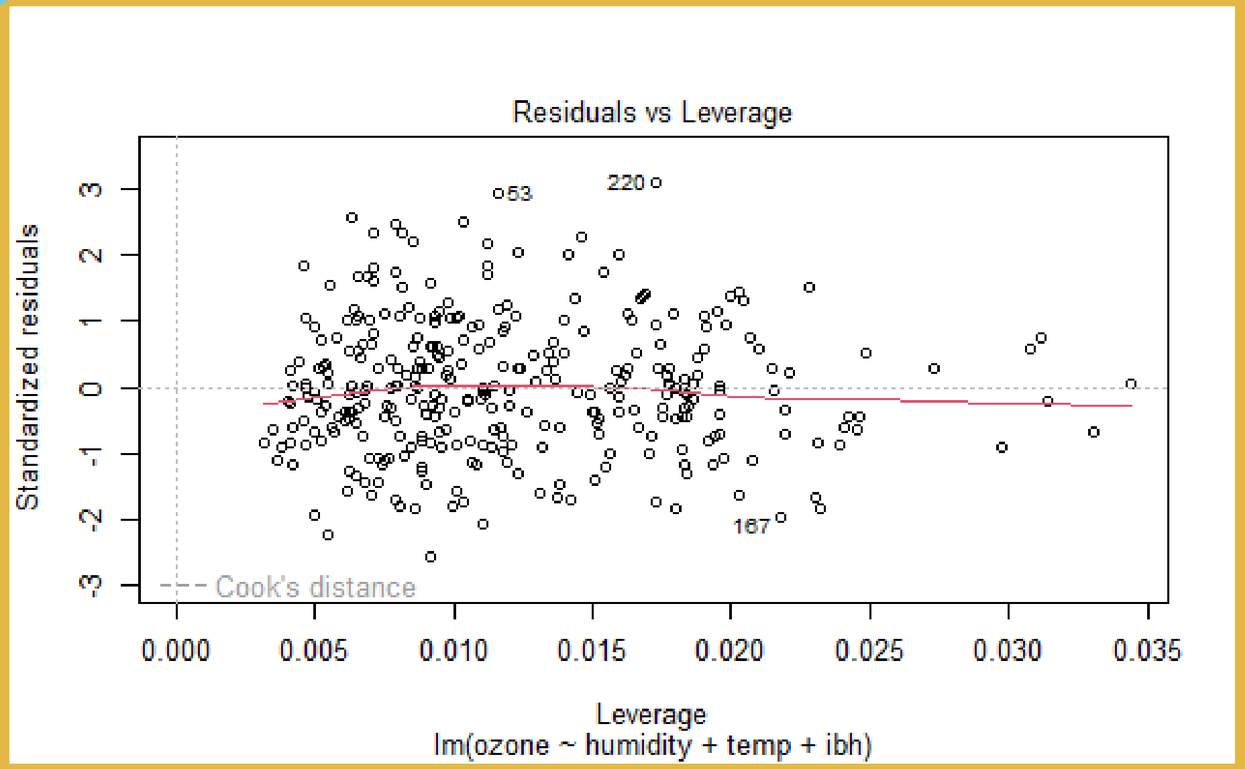
Backwards: 16.603

Stepwise: 16.232

Lasso:16.222

FINAL MODEL

UPLAND MAXIMUM OZONE = Bo+B1{HUMIDITY}+B2{TEMP}+B3{IBH}+E




```
Call:
lm(formula = target ~ temp + ibh + humidity, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9439 -2.8624 -0.1781  2.6414 11.7084

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.109e+01  1.511e+00  -7.337 1.93e-12 ***
temp         3.348e-01  1.981e-02  16.899  < 2e-16 ***
ibh         -9.663e-04  1.519e-04  -6.362 7.13e-10 ***
humidity     7.874e-02  1.228e-02   6.410 5.40e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.075 on 310 degrees of freedom
Multiple R-squared:  0.724,     Adjusted R-squared:  0.7213
F-statistic: 271.1 on 3 and 310 DF,  p-value: < 2.2e-16
R-squared: 0.7240065
Mean Squared Error (MSE): 16.39138
```

R SQUARED: .724

MSE: 16.39

VARIABLES REMOVED: 4

NEXT STEPS AND SUGGESTIONS

CORRECTING LINEAR REGRESSION ASSUMPTIONS

Using Weighted Least Squares Regression, Robust Regression, or Squared methods we can possibly eliminate linearity and homoscedasticity

ADDITIONAL OUTLIER TESTING

More outlier testing may provide clearer insights about why certain variables more heavily violated linear regression assumptions and experiencved multicollinearity