Stats (2122-ST5001) Assignment 5

Alexey Shapovalov (20235952)

19 December, 2021

Introduction

This document is in request to investigate which meteorological factors have had an impact on Winter air pollution in the city of Christchurch. In addition to this, the performance of Government air pollution reduction strategies was evaluated. The report is based on data collected from Christchurch (New Zealand) and contains meteorological and PM10 concentrations readings of the Winter periods (May-August) since the year 2000.

Questions of Interest

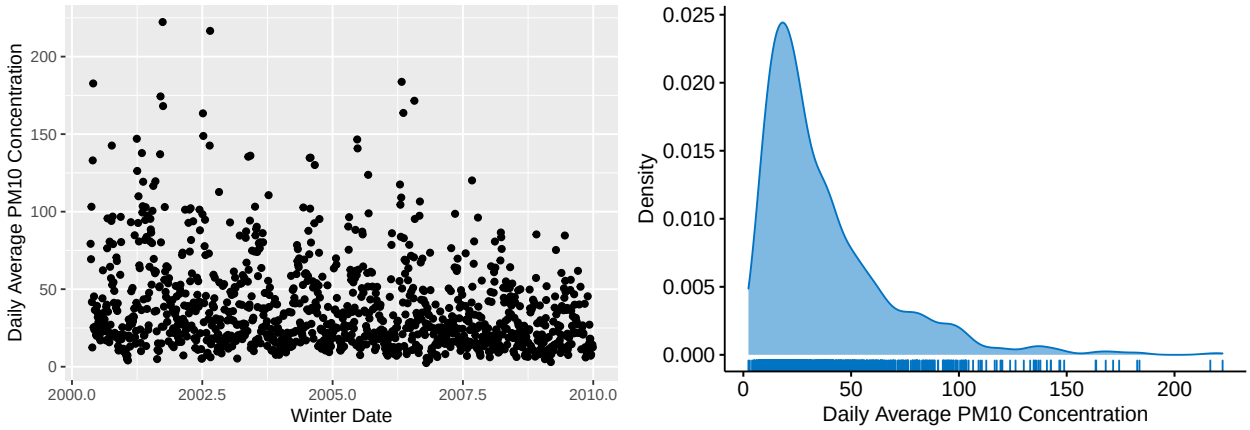
Specifically, the report aims to answer the following questions:

1. **Have the government intervention schemes been effective in reducing the pollution level?** This was evaluated based on an analysis on a line of best fit to the date (feature) and the concentrations of PM10 (dependent variable).
2. **How does the meteorology effect the concentration of PM10?** To answer this, an analysis based on a multiple linear regression model fitted to the meteorological factors (features) and the concentrations of PM10 (dependent variable) was carried out.

Exploratory Analysis

An exploratory investigation was carried out on the data collected. Need summary statistics and another diagram in this section.

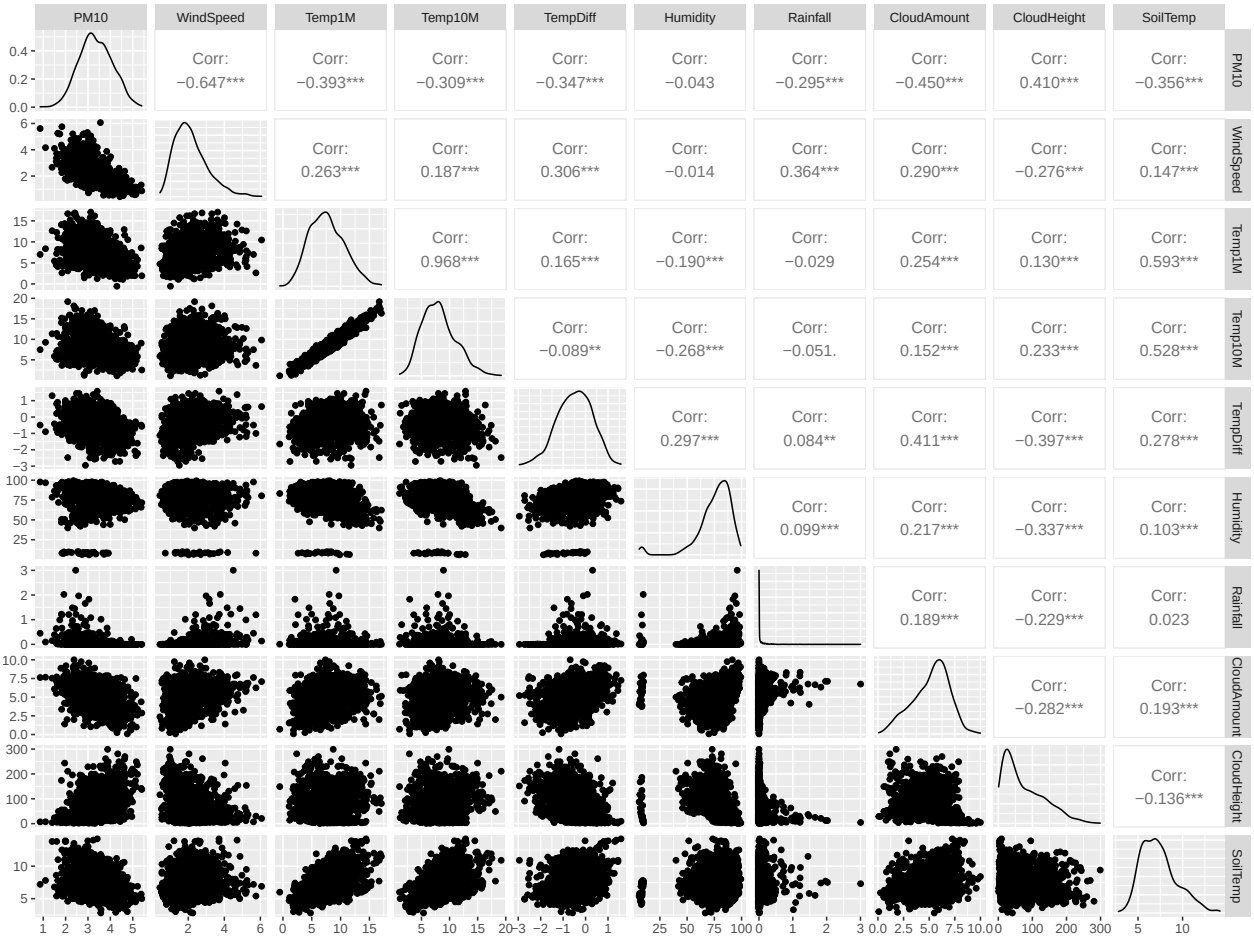
Left: PM10 Over Time, Right: PM10 Density



Observations

- Looking at the average PM10 Concentration values over time it does appear that the values are decreasing as time goes on. This is suggestive that the government polution reduction schemes have had a positive effect.
- The density plot shows that the average values are highly skewed, it makes sense to apply the natural log function for better performance.

Meteorological Factors Correlation Matrix



Observations

- Most variables seem to come from a normal distribution with the exception of Humidity, Rainfall and CloudHeight, these are skewed.
- There is very clear correlation between Temp1M and Temp10M – this makes sense as these are likely to be correlated.
- There does not appear to be much correlation between any two pairs of the other variables.

Formal Analysis

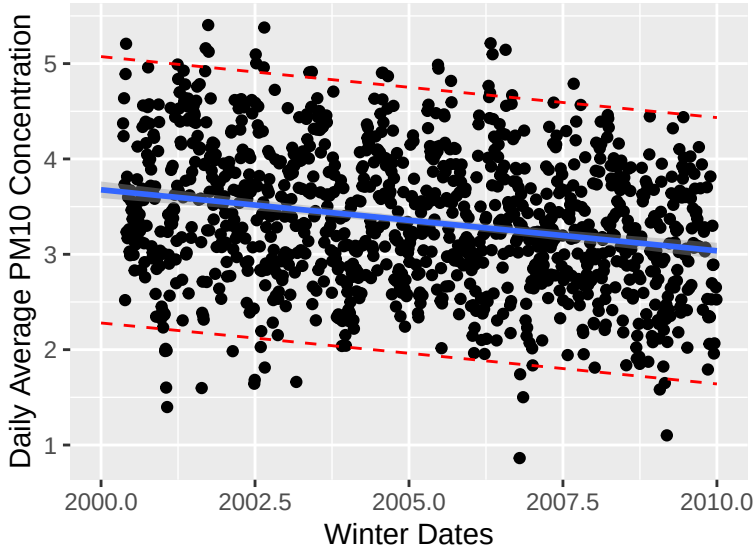
This section contains the outcome of a formal analysis carried out to answer both questions of interest. Due to the skew of the data (see Section PM10 Density plot) all of the models in this report are based on the natural log of PM10.

Effectiveness of Government Intervention Schemes

A simple linear regression model (visualized in Scatterplot with Line of Best Fit and Intervals) was fit to measure the impact of the Government's pollution reduction schemes:

- The model found the intercept to be **131.577** and the coefficient to be **-0.064**. The negative correlation implies that as the time went on concentrations of PM10 decreased.
- The line of best fit resulted in an R-squared score of **0.05836**. This is quite a low score which means the model should not be used for predictive purposes (this is not the goal here).
- The 95/% confidence interval of the coefficient was **-0.079** to **-0.049** – the range is negative and does not cross the 0 mark and as such it is highly likely that the relationship between time and concentrations of PM10 is negative.
- Furthermore, the p-value for the coefficient is $<$ **2.2e-16**. Since this is extremely low (and much below significance level of 0.05) there is enough evidence to reject the null hypothesis; this means that given the data observed it is highly unlikely that there is no relationship between time and concentrations of PM10.

Scatterplot with Line of Best Fit and Intervals



Impact of Meteorology Factors

A multiple regression model was trained using all the meteorological factors as features resulting in the following coefficients and intercept:

Observations

- The f-statistic for the model **187.2** with a p-value of **2.2e-16** signifying that the overall predictive power of the features is good.
- There appears to be collinearity occuring between the features *Temp1M, Temp10M* and *TempDiff*. While all the other features have very low p-value's and 95% confidence intervals that do not cross the 0 mark, with these three it is not the case. Intuitively, it makes sense that there is collinearity between these variables as the temperature 1m above ground is likely to be correlated to the temperature 10m above ground. However, the coefficient of *Temp1M* is positive where as *Temp10M's* coefficient is negative. Similarly, the differential is by definition a function of the two other variables and as such will be collinear with them.
- All of the other features have p-values below the 95% confidence significance level of 0.05.

Table 1: Meteorology Factors Coefficients and Intercepts

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4.770	0.115	41.396	0.000	4.544	4.996
WindSpeed	-0.352	0.018	-19.516	0.000	-0.388	-0.317
Temp1M	-0.041	1.061	-0.038	0.969	-2.123	2.042
Temp10M	-0.005	1.061	-0.005	0.996	-2.088	2.078
TempDiff	0.014	1.059	0.014	0.989	-2.064	2.093
Humidity	0.003	0.001	2.509	0.012	0.001	0.004
Rainfall	-0.169	0.072	-2.339	0.020	-0.310	-0.027
CloudAmount	-0.079	0.009	-8.556	0.000	-0.097	-0.061
CloudHeight	0.003	0.000	11.218	0.000	0.003	0.004
SoilTemp	-0.043	0.009	-4.718	0.000	-0.061	-0.025

Model Selection

For a better interpretation and to reduce the impact of multicollinearity a number of model selection techniques were evaluated:

Best Subset Regression A best subset of 5 was searched and was found to select: *WindSpeed, Temp10M, CloudAmount, CloudHeight, SoilTemp*. With these features the model had an **r-squared** value of **0.5937845** and a **Schwartz's information criterion (BIC)** value of **-987.4477**.

Stepwise Selection Four variants of the stepwise selection search algorithm were carried out. The metric used to evaluate the models was the **Akaike information criterion (AIC)**.

- **Backward Selection**: Starting with each of the features a backward selection removed the *Temp10M and TempDiff* features. This returned the same result as the best set using the BIC metric. The search ran by removing features in this order: 1) Removed *Temp10M*, 2) Removed *TempDiff*.
- **Backward Stepwise Selection**: This search produced the same result as the standard backward selection.
- **Forward Selection**: Starting with just the intercept, the forward selection search found the best set of features to be the same as the best set based on the R squared metric of the exhaustive search. The search ran by adding featurs in this order: 1) Added *WindSpeed*, 2) Added *CloudAmount*, 3) Added *SoilTemp*, 4) Added *CloudHeight*, 5) Added *Temp10M*, 6) Added *Humidity*, 7) Added *Rainfall*
- **Forward Stepwise Selection**: This search produced the same result as the standard forward selection search.

All Possible Regression Since the dataset is relatively small an all possible regression search was carried out, the best feature set was found to be:

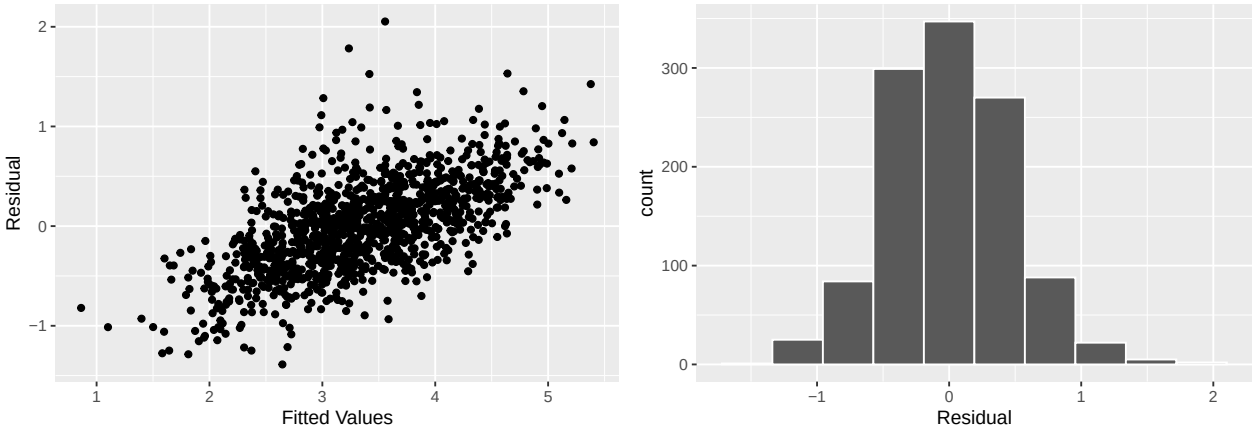
- *WindSpeed, Temp10M, Rainfall, CloudAmount, CloudHeight* based on the best **BIC metric** of **-987.4477**.
- *WindSpeed, Temp1M, Humidity, Rainfall, CloudAmount, CloudHeight, SoilTemp* based on the best **adjusted R squared metric** of **0.5951408**.

Lasso Regression In order to reduce overfitting and as a form of feature selection a lasso regression model was also fit. The lambda hyperparameter was determined based on a **10-fold cross validation search**. The value for lambda that produces the minimum error was **0.001** where as **0.063** was the largest value that kept the error within one standard deviation (1SE) from the minimum. With 1SE the TempDiff variable was effectively removed as its coefficient value was set to 0.

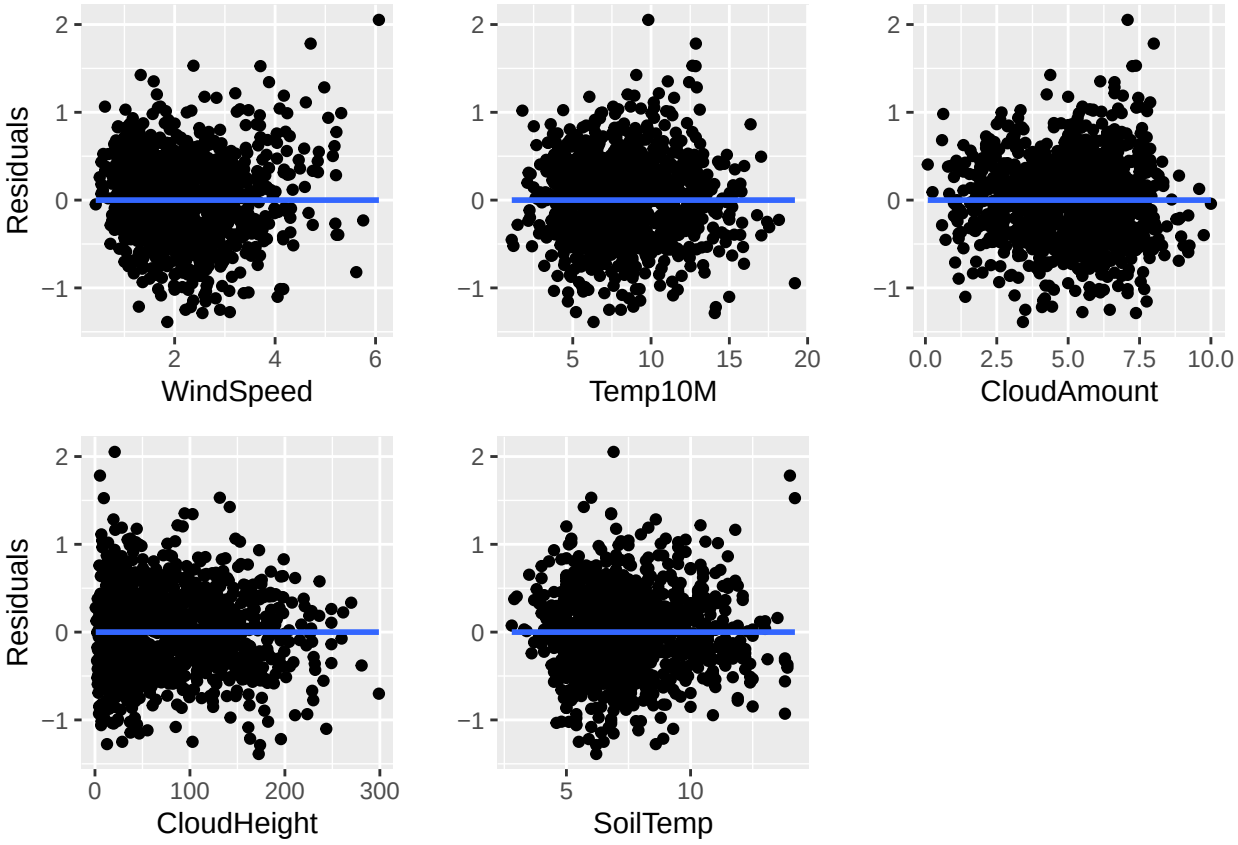
Assumptions Underlying the Model

A number of assumptions are necessarily made when fitting linear models and an attempt was made to validate these. This was done based on the feature subset found from the all possible regressions search.

Left: Residuals vs Fits, Right: Histogram of Residuals



Residuals vs Features



Interpretation

Linearity assumption The assumption is that the population relationship between the mean response and the features is linear. Observed in the residual plot is a trend line; for lower fitted values the residual

tends to be negative where as for higher fitted values the residuals tend to be positive. This is suggestive that the predictions are shifted (biased) towards the average. Linearity between each of the features looks good.

Independence assumption The assumption is that the sample is representative of the population and the subjects are independent. This assumption cannot easily be verified but is still made non the less. However, based on the "Observations" it looks like none of the variables are correlated with each other.

Normality assumption The assumption is that the errors follow a normal distribution, centred about the regression line. It is visible from the histogram of residuals that it the errors follow a normal distribution so this is verified to be true.

Equal spread assumption The assumption is that the variance of the errors is the same for any value of the explanatory variables. There appears to be a change in variance in that the variance seems to be higher in the middle, likely this is because there is a limit on the upper and lower bound of concentrations of PM10.

Final Model

Table 2: Meteorology Factors Coefficients and Intercepts

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	5.013	0.074	67.439	0	4.867	5.158
WindSpeed	-0.373	0.017	-22.552	0	-0.406	-0.341
Temp10M	-0.047	0.006	-7.666	0	-0.059	-0.035
CloudAmount	-0.079	0.009	-9.260	0	-0.096	-0.063
CloudHeight	0.003	0.000	11.421	0	0.003	0.004
SoilTemp	-0.041	0.008	-4.883	0	-0.058	-0.025

Interpretation

- If all the feature variable were equal to zero the predicted value for PM10 concentrations would be between 4.857 and 5.148 with a confidence interval of 95%. This could potentially occur in nature but highly unlikely.
- For each of the features the interpretation is that if all other features are held constant, moving this feature up by a unit would effect the natural log of PM10 concentrations by the correlate amount.
  - Increasing WindSpeed by 1 unit would decrease the natural log of PM10 by 0.373.
  - Increasing Temp10M by 1 unit would decrease the natural log of PM10 by 0.047.
  - Increasing CloudAmount by 1 unit would decrease the natural log of PM10 by 0.079.
  - Increasing CloudHeight by 1 unit would increase the natural log of PM10 by 0.003.
  - Increasing SoilTemp by 1 unit would decrease the natural log of PM10 by 0.041.

Conclusion and Translation

- The formal analysis carried out was able to confirm that the Government air pollution reduction strategies are likely to have had a positive impact as the levels of PM10 trended downwards with time.
- Driving emissions (e.g. colder and wetter days people are more likely to use their solid fuel burners and for longer).
  - As temperature (Temp10M) increases (all things being equal) concentrations of PM10 decrease so it seems colder days are worse on emissions.
  - Rainfall was removed with the all possible regression search so it was not deemed to add any more predictability if it was added to the final set of features.

- Concentrating the emissions (e.g. on calm days the particulates accumulate as they are not dispersed, and temperature inversions lead to insufficient mixing of air layers leading to concentration of the particulates nearer ground level).
  - As wind speed increases (all things being equal) concentrations of PM10 decrease so it seems calm days are more prone to higher emissions.

Appendix: R Code

```
library(GGally)
library(tidyverse)
library(janitor)
library(table1)
library(tolerance)
library(ggplot2)
library(ggridges)
library(viridis)
library(infer)
library(lubridate)
library(moderndive)
library(leaps)
library(ggpubr)
library(kableExtra)
library(glmnet)

set.seed(42)


chch = read.csv("data/chchpollution.csv")

names(chch) # column names

head(chch)

summary(chch)

# WindDirection min is -179.013 and max is 179.5
# This appears to be the angle encoded from the range -180 to 180.
# https://commons.wikimedia.org/wiki/File:Circle_Divided_into_degrees.svg
# As this variable does not actually measure
# on a linear scale it was encoded as a discrete variable.
# https://stackoverflow.com/questions/28327889/
convertToDirection <- function(x) {
  upper <- seq(from = 11.25, by = 22.5, length.out = 17)
  card1 <- c('N', 'NNE', 'NE', 'ENE', 'E', 'ESE', 'SE',
             'SSE', 'S', 'SSW', 'SW', 'WSW',
             'W', 'WNW', 'NW', 'NNW', 'N')
  ifelse(x>360 | x<0,NA,card1[findInterval(x,upper,rightmost.closed = T)+1])
}
chch$WindDirection = chch$WindDirection + 180  # Make range 0 - 360
chch$WindDirection = as.factor(sapply(chch$WindDirection, convertToDirection))

head(chch)
```

```r
summary(chch %>% select(Rainfall))

# Make date column
chch = chch %>% mutate(date = dmy(paste(Day, Month, Year, sep = "/")))


# Data wrangling:
# Extract the winter data and from year 2000,
# as government interventions started in 2002
# There are 123 days of winter, so create
# date index over winters days (not calendar days)
# To make your analysis easier, the days with
# a missing PM10 measurement are also ignored
chchwinter = chch %>%
  filter(Month %in% 5:8, Year >= 2000) %>%
  mutate(dateindex = 2000 + (row_number() - 1)/123) %>%
  filter(!is.na(PM10))
chchwinter = na.omit(chchwinter)

PM10OverTime <- ggplot(chchwinter, aes(x = dateindex, y = PM10)) +
  geom_point() +
  xlab("Winter Date") + ylab("Daily Average PM10 Concentration")
density <- ggdensity(
  chchwinter,
  x='PM10',
  fill = "#0073C2FF",
  color = "#0073C2FF",
  rug = TRUE,
  xlab = "Daily Average PM10 Concentration",
  ylab = "Density"
)
gridExtra::grid.arrange(PM10OverTime, density, ncol=2)


chchwinterUpdated <- chchwinter
chchwinterUpdated$PM10 <- log(chchwinter$PM10)
ggdensity(
  chchwinterUpdated,
  x='PM10',
  fill = "#0073C2FF",
  color = "#0073C2FF",
  rug = TRUE,
  xlab = "Daily Average PM10 Concentration",
  ylab = "Density"
)

# Use natural log of PM10 to reduce skew
chchwinter$PM10 <- log(chchwinter$PM10)


columns <- c(
  "PM10",
  "WindSpeed",
```

```
  "Temp1M",
  "Temp10M",
  "TempDiff",
  "Humidity",
  "Rainfall",
  "CloudAmount",
  "CloudHeight",
  "SoilTemp"
)
chchwinter %>% select(columns) %>% ggpairs(progress = FALSE)


model <- lm(PM10 ~ dateindex, data=chchwinter)
get_regression_table(model)
summary(model)


PM10.range <- data.frame(dateindex = seq(2000, 2010, 0.1))
pred.int <- predict(model, newdata=PM10.range, interval="prediction")
PM10.range = cbind(PM10.range, pred.int)

chchwinter %>%
  ggplot(aes(x = dateindex, y = PM10)) +
  geom_point() +
  stat_smooth(method = lm, fullrange = TRUE) +
  geom_line(data = PM10.range, aes(y = lwr), color = "red", linetype = "dashed") +
  geom_line(data = PM10.range, aes(y = upr), color = "red", linetype = "dashed") +
  labs(x = "Winter Dates", y = "Daily Average PM10 Concentration")



chchwinter.formula = PM10 ~
  WindSpeed +
  Temp1M +
  Temp10M +
  TempDiff +
  Humidity +
  Rainfall +
  CloudAmount +
  CloudHeight +
  SoilTemp
chchwinter.model = lm(chchwinter.formula, data = chchwinter)

df <- get_regression_table(model = chchwinter.model)
kbl(df, caption = "Meteorology Factors Coefficients and Intercepts", booktabs = T) %>%
kable_styling(latex_options = c("striped", "hold_position"))


summary(chchwinter.model)


models = regsubsets(chchwinter.formula, data=chchwinter, nvmax=5)
summary(models)
```

```
summary(models)$bic
summary(models)$rsq


step(chchwinter.model, direction="backward")


step(chchwinter.model, direction="both")


min.model <- lm(PM10 ~ 1, data=chchwinter)
step(
  min.model,
  direction="forward",
  scope=list(
    lower=~ 1,
    upper=~
      WindSpeed +
      Temp1M +
      Temp10M +
      TempDiff +
      Humidity +
      Rainfall +
      CloudAmount +
      CloudHeight +
      SoilTemp
  )
)


step(
  min.model,
  direction="both",
  scope=list(
    lower= ~ 1,
    upper=~
      WindSpeed +
      Temp1M +
      Temp10M +
      TempDiff +
      Humidity +
      Rainfall +
      CloudAmount +
      CloudHeight +
      SoilTemp
  )
)


models = regsubsets(chchwinter.formula, data=chchwinter)
summary(models)
summary(models)$bic
summary(models)$adjr2
```

```
chchwinter.bestcols.formula = PM10 ~
  WindSpeed +
  Temp10M +
  CloudAmount +
  CloudHeight +
  SoilTemp
chchwinter.bestcols.model = lm(chchwinter.bestcols.formula, data=chchwinter)


X = model.matrix(chchwinter.formula, chchwinter)[, -1]
y = chchwinter$PM10
fit.cvlasso = cv.glmnet(X, y, alpha = 1, lambda = 10^seq(-3, 6, 0.1))
plot(fit.cvlasso)
# value of lambda that gives minimum cvm
fit.cvlasso$lambda.min
# largest value of lambda such that error
# is within 1 standard error of the minimum
fit.cvlasso$lambda.1se


coeffdf <- cbind(coefficients(
  fit.cvlasso,
  s = "lambda.min")[-1],
  coefficients(chchwinter.model)[-1]
)
colnames(coeffdf) <- c('Lasso (1SE)', 'Maximal')
kbl(
  coeffdf,
  caption = "Lasso Coefficients compared to Maximal Regression",
  booktabs = T
) %>%
kable_styling(latex_options = c("striped", "hold_position"))


residualVsFits <- get_regression_points(chchwinter.bestcols.model) %>%
  ggplot(aes(x = PM10, y = residual)) +
  geom_point() +
  labs(x = "Fitted Values", y = "Residual")
histogram <- get_regression_points(chchwinter.bestcols.model) %>%
  ggplot(aes(x = residual)) +
  geom_histogram(color = "white", bins = 10) +
  labs(x = "Residual")


gridExtra::grid.arrange(residualVsFits, histogram, ncol = 2)
res = cbind(chchwinter, resids = residuals(chchwinter.bestcols.model))

plot1 = ggplot(res, aes(x = WindSpeed, y = resids)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE) +
    labs(x = "WindSpeed", y = "Residuals")
plot2 = ggplot(res, aes(x = Temp10M, y = resids)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE) +
```

```
    labs(x = "Temp10M", y = "")
plot3 = ggplot(res, aes(x = CloudAmount, y = resids)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE) +
    labs(x = "CloudAmount", y = "")
plot4 = ggplot(res, aes(x = CloudHeight, y = resids)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE) +
    labs(x = "CloudHeight", y = "Residuals")
plot5 = ggplot(res, aes(x = SoilTemp, y = resids)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE) +
    labs(x = "SoilTemp", y = "")

gridExtra::grid.arrange(plot1, plot2, plot3, plot4, plot5, ncol = 3)


df <- get_regression_table(model = chchwinter.bestcols.model)
kbl(df, caption = "Meteorology Factors Coefficients and Intercepts", booktabs = T) %>%
kable_styling(latex_options = c("striped", "hold_position"))


summary(chchwinter.model)
```