Predictive Data Analysis

Weather in Australia

Jennie Wang

A Description of Data

I select the data from Albury weather station in 2008-2009 to test my prediction of Australia weather. I choose this dataset because the weather is a natural phenomenon that hard to be manipulated by policy or other artificial influences that may influence data precision. Weather can be influenced in many aspects so that different variables have direct relationships.

The data is from Kaggle. (1)

> ***KAGGLE*** *is an online community of data scientists and machine learners, owned by Google LLC. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. (2)*
>
> *This dataset contains about 10 years of daily weather observations from numerous Australian weather stations. Observations were drawn from numerous weather stations. The daily observations are available from http://www.bom.gov.au/climate/data. Copyright Commonwealth of Australia 2010, Bureau of Meteorology. (1)*

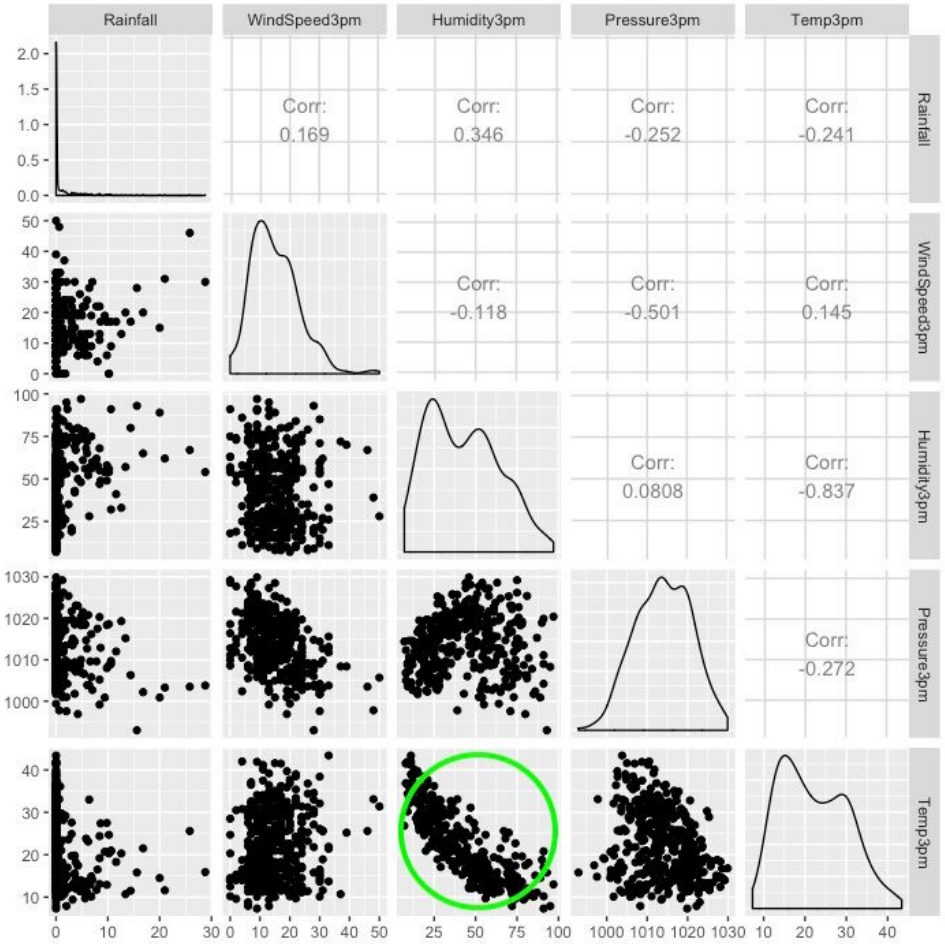
This dataset has 24 variables/columns. To make a prediction, I only need two variables so I clear the data in advance by choosing 5 columns. My variables include ‘Rainfall’, ‘WindSpeed3pm’, ‘Humidity3pm’, ‘Pressure3pm’, ‘Temp3pm’. The followings are the meaning of each variable:

> *Rainfall / The amount of rainfall recorded for the day in mm; WindSpeed3pm / Wind speed (km/hr) averaged over 10 minutes prior to 3pm; Humidity3pm / Humidity (percent) at 3pm; Pressure3pm / Atmospheric pressure (hpa) reduced to mean sea level at 3 pm; Temp3pm / Temperature (degrees C) at 3pm*

There are two limitations to this dataset. First, there is an “NA” in the record of rainfall on 12/16/2008. To help RStudio run data smoothly, I remove this row. The other limitation is only Albury station data starts from 2008. Other stations’ data all starts from 2009 to 2017. Then, when I finished my prediction of 2008-2009 data in Albury, I can’t use data from another station to make a comparison with the data in the same year range.

Justification for Support

In order to find the two correlation in five variables, I use packages: "GGally" and select a total of five columns to make a comparison visually. Three charts come up and bright my eyes, WindSpeed3pm & Pressure3pm, Humidity3pm & Temp3pm, Pressure3pm & Temp3pm. I choose the second one to make a prediction because it has the strongest trend in visual which is better for the regression model. From the correlation data on the upper side, I find Humidity3pm & Rainfall has the highest correlation number: 0.346 while Temp3pm & Humidity3pm is -0.837. However, correlation doesn't mean causation so that I still choose Humidity3pm & Temp3pm to my " Predictive Data Analysis".



I found an article about the relationship between temperature and humidity.

> *Temperature affects humidity, which in turn affects the potential for precipitation. Relative humidity represents a percentage of water vapor in the air that changes when the air temperature changes. For example, a completely saturated parcel of air at constant pressure cannot hold any more water molecules, giving it a relative humidity of 100 percent. (5)*

From this perspective, the temperature is my independent variable while the humidity is my dependent variable. This definition can help me make my chart in Rstudio to explore more information between those two variables.

I use the Linear model to minimizing some of the squares to create a line that best fits my data. I set Humidity3pm as my dependent variable and Temp3pm as my independent variable. Coefficients: 91.869(Intercept), Temp3pm: -2.218. This tells me when the temperature is -2.218 degrees the humidity (percent) is 91.869. In order to see more information, I use the 'summary' to see the data includes residuals and more coefficients information. Plus, multiple R-squared is a measurement that how closely this model fits my data and I get 0.7004 which is not bad.

The correlation between 2 variables

I use different formulas and packages in RStudio to explore the correlation between Humidity3pm and Temp3pm. The fitted model helps me to see how Rstudio uses it to predict 390 data and each point is on the pattern.

> *The values for an output variable that have been predicted by a **model fitted** to a set of data. ... **Fitted** values are generated by extending the **model** past known data points in order to predict unknown values. Also called predicted values. (6)*

```
360
35.0798606 41.5130488 36.6326991 20.8824796 35.9671969 37.0763673 34.8580265 45.9497304 39.5165421 41.734
8829
      361        362        363        364        365        366        367        368        369
370
35.3016947 25.7628293 32.4178516 34.8580265 28.2030042 23.9881567 43.5095555 39.9602103 35.0798606 45.284
2282
      371        372        373        374        375        376        377        378        379
380
34.1925243 27.7593360 23.9881567 19.7733092 10.0126098 22.2134840 36.6326991 30.4213449 24.4318248 19.329
6410
      381        382        383        384        385        386        387        388        389
390
12.2309505  8.6816053 24.4318248 30.8650131 26.6501656 41.5130488 16.2239639 16.2239639 14.0056232 22.878
9863
```

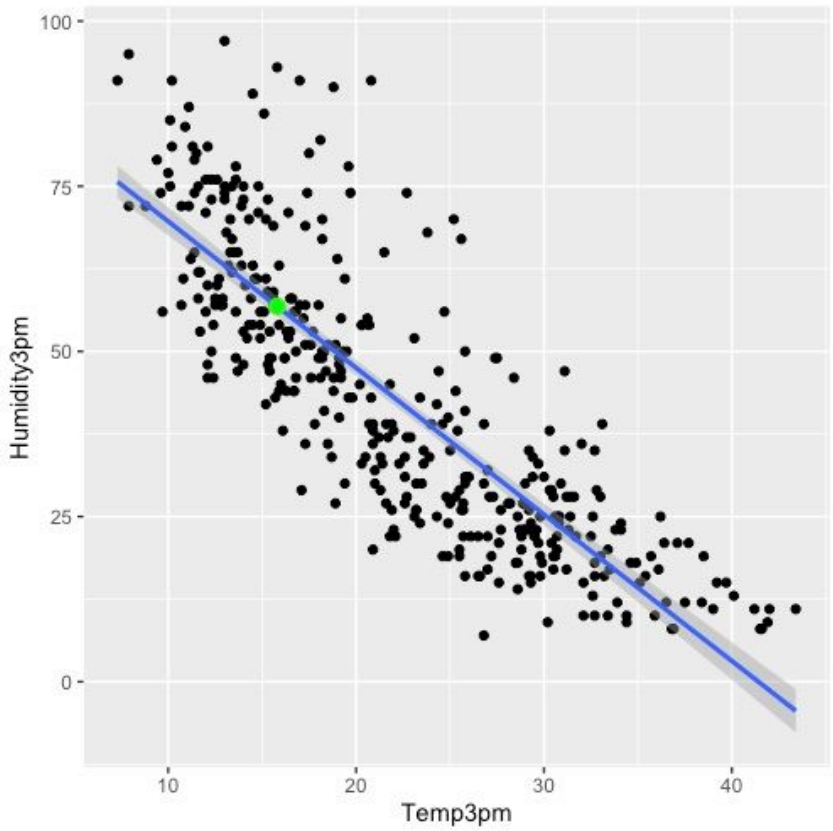
From residuals, I know how much my model of from I actually measured. How much the measured value from what was predicted by the line. Where my model is really far off. Broom packages help me to see how far the predicted model from the original dataset. From 'lm_matrix' formula, I know my best residual is 45.3. I can also see my standard error from the data to know how much outside of the regression line is counted by this model.

> *A residual is a vertical distance between a data point and the regression line. Each data point has one residual. They are positive if they are above the regression line and negative if they are below the regression line. If the regression line actually passes through the point, the residual at that point is zero. (7)*

```
Residuals:
    Min      1Q  Median      3Q     Max
-25.506  -7.743  -1.876   6.654  45.272

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 91.86938    1.73956   52.81   <2e-16 ***
Temp3pm     -2.21834    0.07365  -30.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

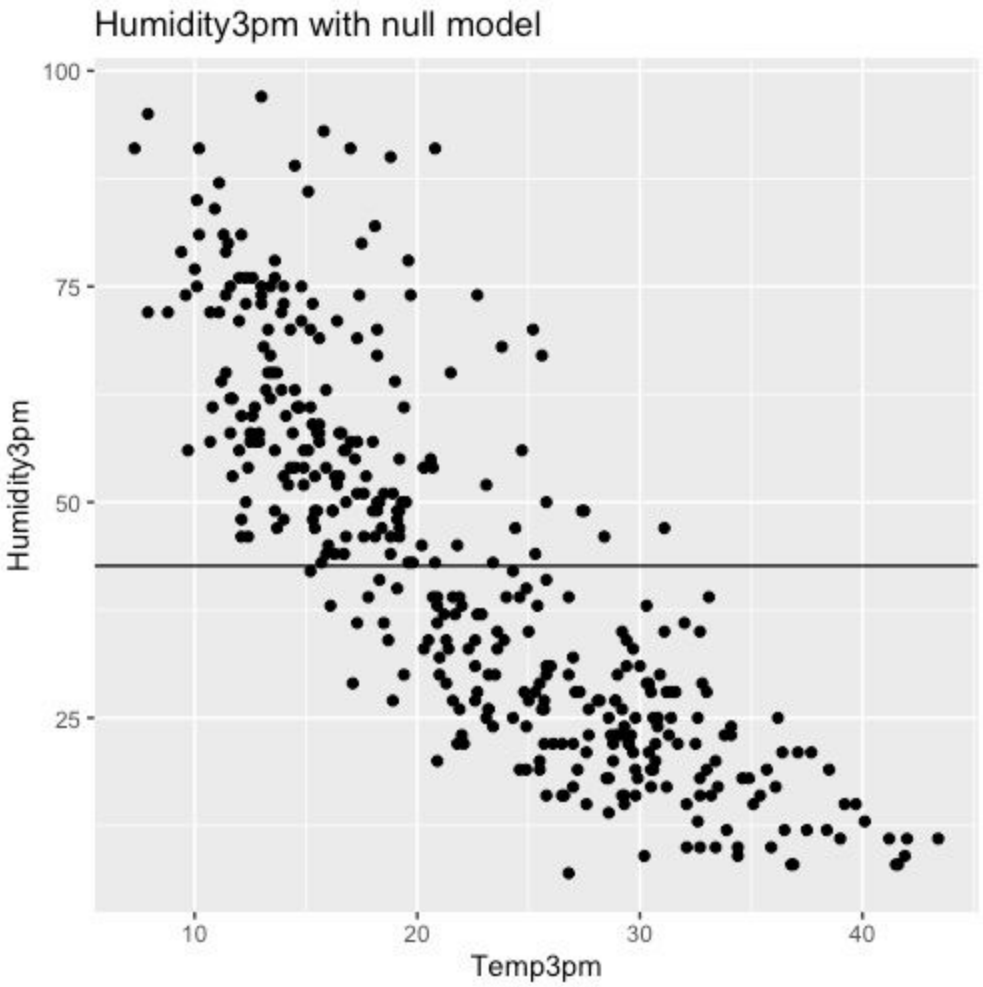
Visually, the grey area contains points that are calculated by Rstudio while others outside points are not be calculated to make a linear model. The green point is my fitted value with my predicted new data so it is on the line directly.



From my predict model, when my Temp3pm is 15.8, my fitted value of new humanity is 56.8196. I type the mean formula in Rstudio to get the mean value which is 42.62051. Using a null-model to make a comparison with my predict linear model.

Null model(4) Coefficients: (Intercept) 42.62

> *In mathematics, for example in the study of statistical properties of graphs, a* ***null model*** *is the type of random object that matches one specific object in some of its features, or more generally satisfies a collection of constraints, but which is otherwise taken to be an unbiasedly random structure. The null model is used as a term of comparison, to verify whether the object in question displays some non-trivial features (properties that wouldn't be expected on the basis of chance alone or as a consequence of the constraints), such as community structure in graphs.(4)*


Humidity3pm with null model
100
75
50
25
Humidity3pm
10
20
30
40
Temp3pm

The Explanation of Correlation

Temperature & humidity has a strong correlation and their correlation is Negative. When one value decreases as the other increases can be called Negative correlation.

> *Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. For example, height and weight are related; taller people tend to be heavier than shorter people. The relationship isn't perfect. People of the same height vary in weight, and you can easily think of two people you know where the shorter one is heavier than the taller one. Nonetheless, the average weight of people 5'5" is less than the average weight of people 5'6", and their average weight is less than that of people 5'7", etc. Correlation can tell you just how much of the variation in peoples' weights is related to their heights. (3)*

Higher R-squared values represent smaller differences between the observed data and the fitted values. In general, the higher the R-squared, the better the model fits your data. My R-squared value is 0.7004 which means my model isn't far from the original dataset.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 91.86938    1.73956   52.81   <2e-16 ***
Temp3pm     -2.21834    0.07365  -30.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.73 on 388 degrees of freedom
Multiple R-squared:  0.7004,    Adjusted R-squared:  0.6997
F-statistic: 907.3 on 1 and 388 DF,  p-value: < 2.2e-16
```

Appendix

```
install.packages("ggthemes")
#install.packages("dplyr")
install.packages("broom")
install.packages("GGally")

library(tidyverse)
library(lubridate)
library(broom)
library(GGally)

ggpairs(data = weatherAUS, columns= 1:5)

weatherAUS <- read.csv("Desktop/weatherAUS.csv", header=TRUE, check.names = FALSE)

glimpse(weatherAUS)

ggplot(weatherAUS, aes(x=weatherAUS$Temp3pm, y=weatherAUS$Humidity3pm)) +
  geom_point() +labs(title = "weatherAUS Humidity3pm by Temp3pm")

ggplot(weatherAUS, aes(x=weatherAUS$Temp3pm, y=weatherAUS$Humidity3pm)) +
  geom_point() +
  stat_smooth(method = "lm", se=FALSE) +
  labs(title = "weatherAUS WindSpeed3pm by Temp3pm")

lm_weatherAUS <- lm(Humidity3pm ~ Temp3pm, data = weatherAUS)
lm_weatherAUS

summary(lm_weatherAUS)

coef(lm_weatherAUS)
```

```
fitted_weatherAUS <- fitted.values(lm_weatherAUS)
fitted_weatherAUS

residual_weatherAUS <- residuals(lm_weatherAUS)
residual_weatherAUS

lm_matrix_weatherAUS <- broom::augment(lm_weatherAUS)
head(lm_matrix_weatherAUS)

lm_matrix_weatherAUS %>%
  arrange(desc(.resid)) %>%
  head()

lm_matrix_weatherAUS$.resid_abs <- abs(lm_matrix_weatherAUS$.resid)
lm_matrix_weatherAUS %>%
  arrange(desc(.resid_abs)) %>%
  head()

weatherAUS %>%
  filter(Temp3pm == 20.8)

new_weatherAUS <- data.frame("Temp3pm" = 15.8, 17.0)
predict(lm_weatherAUS, newdata = new_weatherAUS)

myweatherAUS <- broom::augment(lm_weatherAUS, newdata=new_weatherAUS)
myweatherAUS

ggplot(data = weatherAUS, aes(x=Temp3pm, y=Humidity3pm))+
  geom_point() +
  stat_smooth(method = "lm") +
  geom_point(data = myweatherAUS, aes(y=.fitted), size =3, color="Green") +
  labs("weatherAUS Humidity3pm by Temp3pm")
```

```
weatherAUS_null <- lm(Humidity3pm ~ 1, data=weatherAUS)
weatherAUS_null

mean(weatherAUS$Humidity3pm)

ggplot(data = weatherAUS, aes(x=Temp3pm, y=Humidity3pm)) +
  geom_point() +
  stat_smooth(method = "lm") +
  geom_hline(yintercept = 42.62) +
  labs(title = "Humidity3pm with null model")

summary(lm_weatherAUS)
```

References

(1) Source & Acknowledgements, https://www.kaggle.com/jsphyg/weather-dataset-rattle-package

(2) Kaggle, from Wikipedia, the free encyclopedia https://en.wikipedia.org/wiki/Kaggle

(3) Correlation, https://www.surveysystem.com/correlation.htm

(4) Null_model, https://en.wikipedia.org/wiki/Null_model

(5) Dotson, J. Dianne, *How Temperature & Humidity are Related* https://sciencing.com/temperature-ampamp-humidity-related-7245642.html

(6) Fitted value, http://www.businessdictionary.com/definition/fitted-value.html

(7) Residual, https://www.statisticshowto.datasciencecentral.com/residual/