HUMIDITY AND TEMP

By Steven Tucker

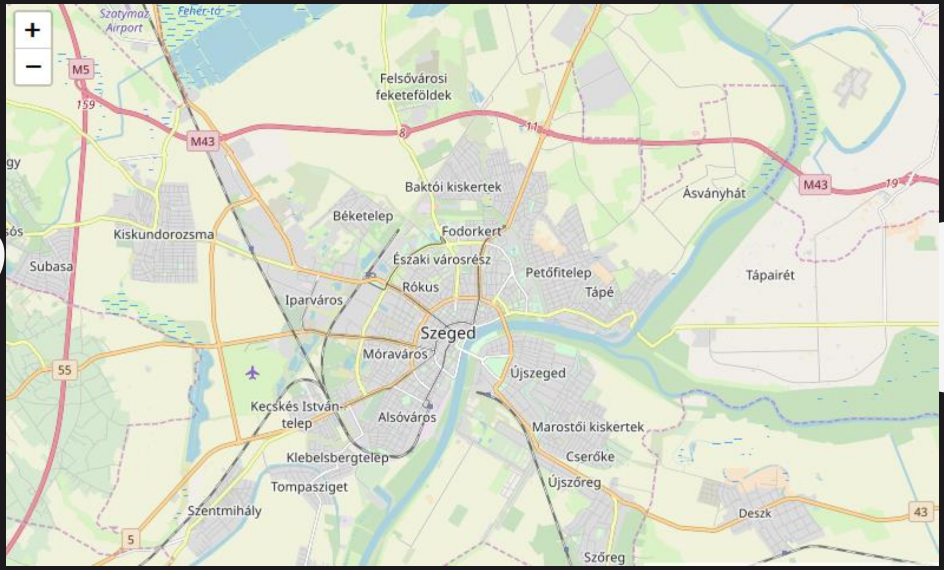
HUMIDITY AND TEMP

The data originates from Szegede, Hungary where they collected weather data once an hour every day for 10 consecutive years. Collecting so much data without lapse in any of the measurements is remarkable in itself. The whole data set contains several columns I am electing to ignore for the purpose of this project. For this project I centered my focus around the Humidity levels and what relationship if any it held with Temperature and Apparent Temperature. Furthermore I wanted to know if the Apparent Temperature was more affected by humidity than the real temperature and if so, in what way? The independent variable for these exercises was Humidity, dependent variables were the temperatures and the Control test was the relationship between Humidity and real temperature.

HOW DATA WAS COLLECTED AND THE BACK STORY

Unfortunately I was not able to trace the origin of the csv file to find out the reason they collected the data, or for that matter how. I would assume the data was collected with the regular tools used to forecast weather for the area. As for the story surrounding the data, I actually looked into the city that the data was gathered in and it is called the “City of Sunshine”, which was an interesting label for a city to earn. According to Wikipedia, the city sees a higher average number of daylight hours and I think this may have contributed to the reason data was collected on the city.

HOW DATA WAS COLLECTED AND THE BACK STORY



Map of Szegede showing surrounding land and river

This snippet of average mean sunshine hours (bottom row) expands on the fact I shared in the previous slide.

Climate data for Szeged (1971-2000)													[hide]
Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
Average high °C (°F)	2.8 (37.0)	5.7 (42.3)	11.6 (52.9)	16.9 (62.4)	22.4 (72.3)	25.5 (77.9)	27.7 (81.9)	27.6 (81.7)	23.3 (73.9)	17.2 (63.0)	8.9 (48.0)	4.1 (39.4)	16.1 (61.1)
Daily mean °C (°F)	−0.8 (30.6)	1.2 (34.2)	5.9 (42.6)	10.8 (51.4)	16.3 (61.3)	19.2 (66.6)	20.8 (69.4)	20.8 (69.4)	16.4 (61.5)	11.0 (51.8)	4.7 (40.5)	0.9 (33.6)	10.6 (51.1)
Average low °C (°F)	−3.8 (25.2)	−2.6 (27.3)	0.5 (32.9)	5.2 (41.4)	10.3 (50.5)	13.0 (55.4)	14.3 (57.7)	14.0 (57.2)	10.3 (50.5)	5.6 (42.1)	1.2 (34.2)	−2.0 (28.4)	5.5 (41.9)
Average precipitation mm (inches)	24 (0.9)	23 (0.9)	25 (1.0)	40 (1.6)	51 (2.0)	68 (2.7)	53 (2.1)	56 (2.2)	37 (1.5)	35 (1.4)	38 (1.5)	39 (1.5)	489 (19.3)
Mean monthly sunshine hours	59	94	143	173	234	252	278	263	199	153	77	53	1,978

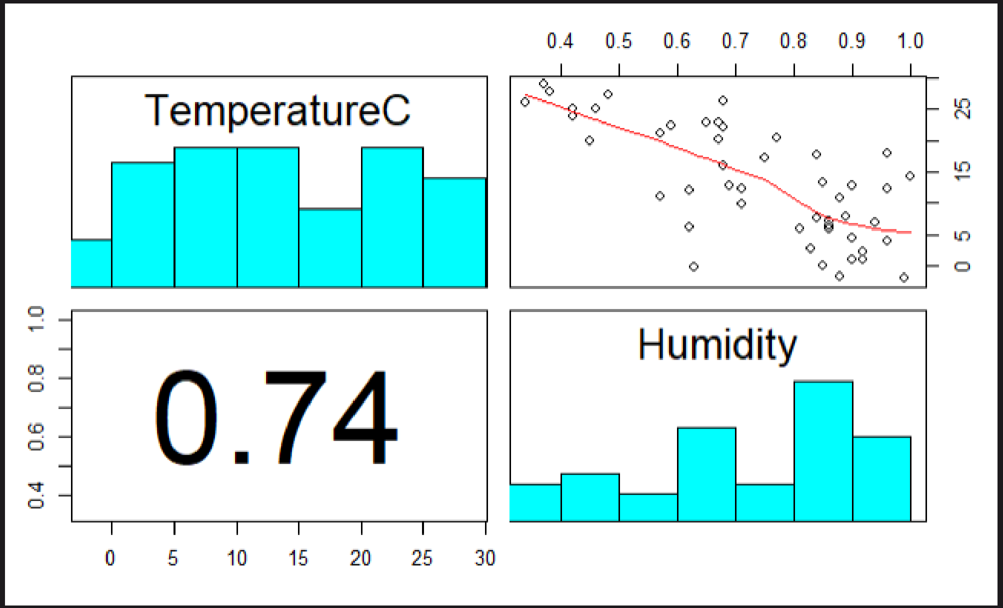
THE RESEARCH QUESTION

My research question is- "How does humidity change the Apparent Temperature?"

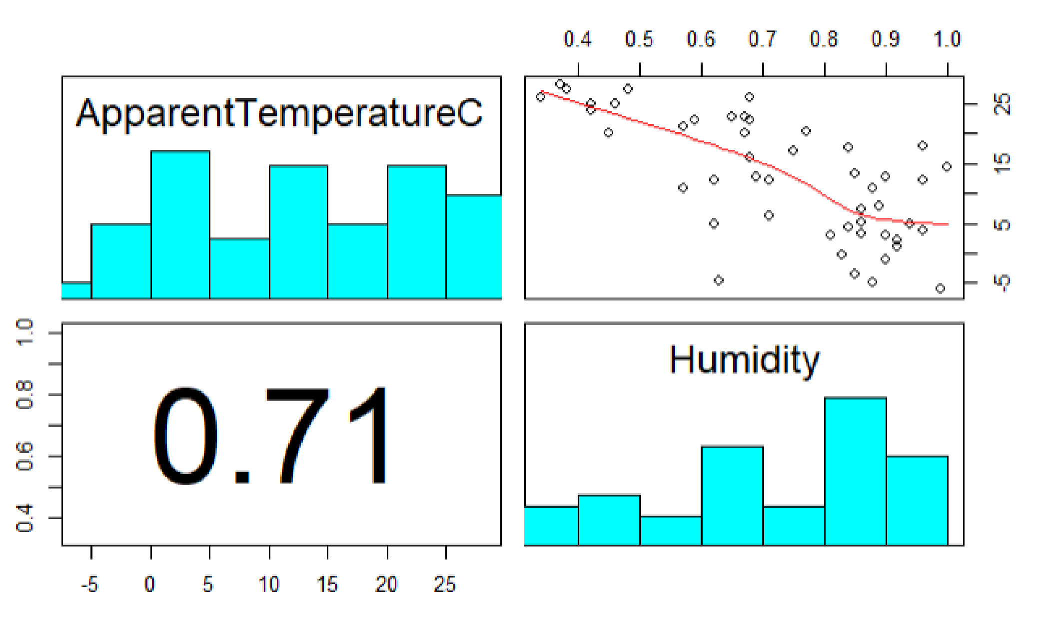
ASSUMPTIONS

- The trend is linear.
- The additional assumptions we need to check can be summarized in the following formula
- $\epsilon_i \sim N\ (0, \sigma^2)$
- This assumes also that the errors are independent.

TRENDS



TemperatureC
Humidity
0.74
0.4 0.5 0.6 0.7 0.8 0.9 1.0
0 5 10 15 20 25 30
0 5 15 25
0.4 0.6 0.8 1.0

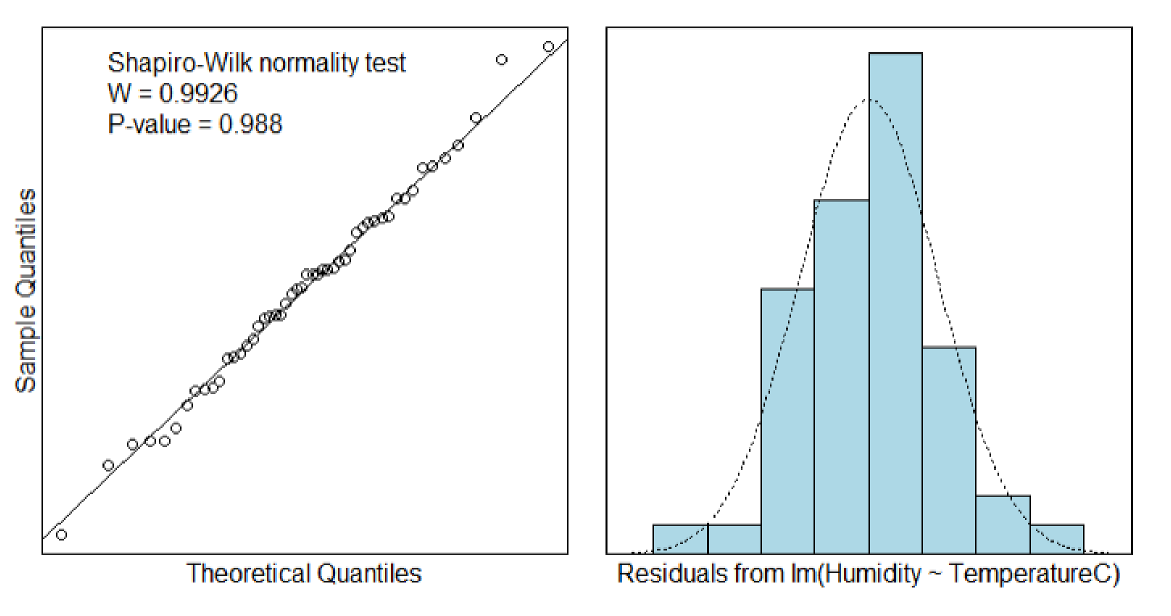


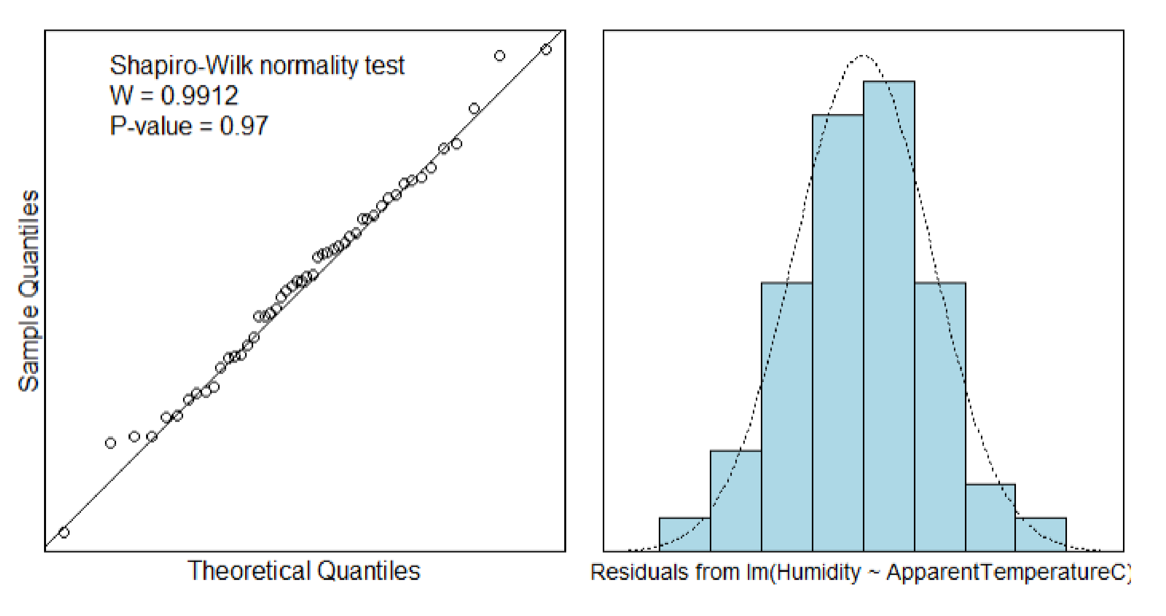
ApparentTemperatureC
Humidity
0.71
0.4 0.5 0.6 0.7 0.8 0.9 1.0
-5 0 5 10 15 20 25
-5 5 15 25
0.4 0.6 0.8 1.0

INDEPENDENT

The collection of one measurement in no way impacted another measurement, all information was gathered independently.

$$N\,(0, \sigma^2)$$


Shapiro-Wilk normality test
W = 0.9926
P-value = 0.988
Sample Quantiles
Theoretical Quantiles
Residuals from lm(Humidity ~ TemperatureC)


Shapiro-Wilk normality test
W = 0.9912
P-value = 0.97
Sample Quantiles
Theoretical Quantiles
Residuals from lm(Humidity ~ ApparentTemperatureC)

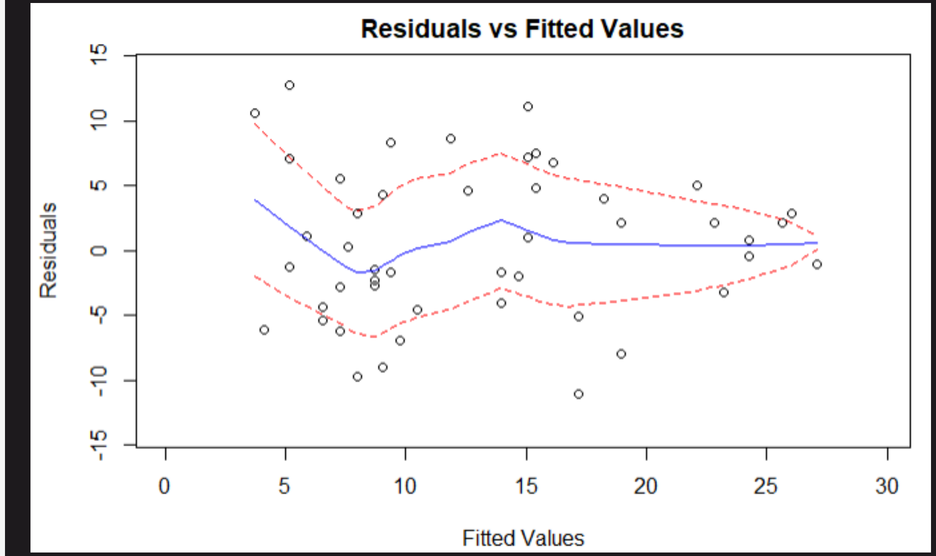
JUSTIFYING NORMALITY

The claim that the data comes from a normal distribution is supported by the Shapiro-Wilkes test. In both cases the plots appear to be normal.

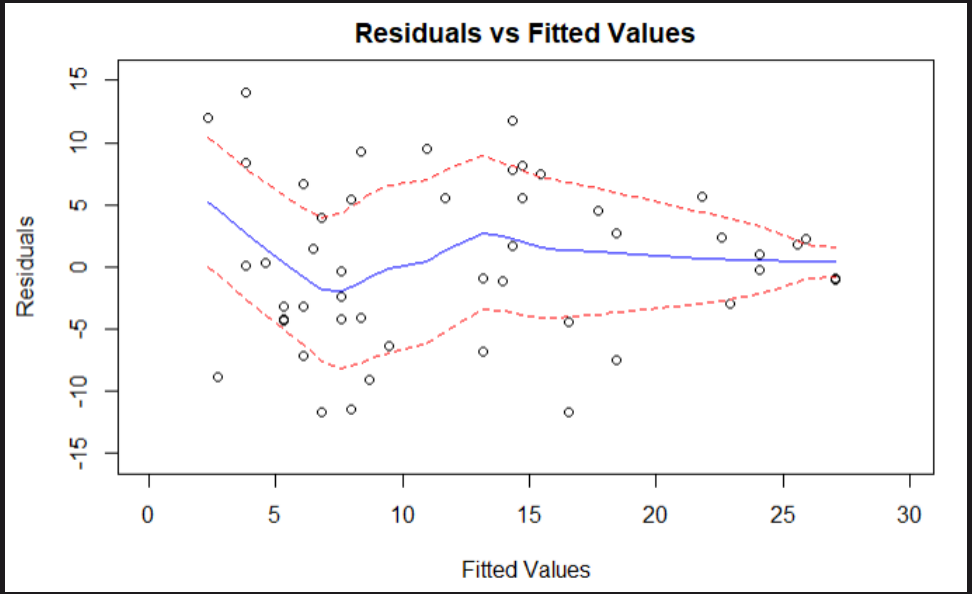
The NULL Hypothesis given by the test is $H_0\!:\ \in_i\ \sim N$ with a p-val of .988 and .97, both of which are far too high to reject the NULL at the 0.05 level of significance.

Since n>30 we could use CLT to validate the assumption that $\bar{Y}\sim N$ and we conclude the normality assumption was justified

CONSTANT VARIANCE ABOUT LINE



Control-Real Temp v Humidity



Experiment-Apparent Temp v Humidity

RESIDUALS

The residuals are mostly centered around the origin

Overall impression is that the residuals are consistent with $\in_i \sim N(0, \sigma^2)$

THE LINEAR MODEL

Used SLR model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Estimated by

$$y_i = \widehat{\beta_0} + \widehat{\beta_1} x_i + r_i$$

Where $\widehat{\beta_i}$ is estimated using the least squares

$$\widehat{\beta_1} = \frac{SS_{xy}}{SS_{xx}}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

ASSESSING THE MODEL

My variability is wide ranging so prediction is nearly out of the question. My multiple R-squared is just above 50% so about 50% of the variation in distance is explained through the Humidity which does not make this model a great predictor.

The research question did not look to predict however, so the question still needs to be assessed by the line of best fit.

LEAST SQUARE ESTIMATES

The research question is one that analyzes the slopes of Temperature and Apparent Temperature as Humidity increases. These slopes are given in the Least Square Estimates shown below-

```
Call:
lm(formula = TemperatureC ~ Humidity + I(Humidity^2), data = weather.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9046 -1.6020  0.2689  1.1766  3.3584 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)      81.49      37.36   2.181   0.0342 *
Humidity       -168.31      88.91  -1.893   0.0645 .
I(Humidity^2)    90.63      52.66   1.721   0.0918 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.846 on 47 degrees of freedom
Multiple R-squared:  0.2808,	Adjusted R-squared:  0.2502 
F-statistic: 9.177 on 2 and 47 DF,  p-value: 0.000432
```

```
Call:
lm(formula = ApparentTemperatureC ~ Humidity + I(Humidity^2), 
    data = weather.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9768 -2.2010  0.3096  1.3638  7.0555 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      81.14      52.81   1.537    0.131
Humidity       -179.35     125.66  -1.427    0.160
I(Humidity^2)    98.82      74.43   1.328    0.191

Residual standard error: 2.609 on 47 degrees of freedom
Multiple R-squared:  0.1288,	Adjusted R-squared:  0.09175 
F-statistic: 3.475 on 2 and 47 DF,  p-value: 0.03913
```

ANSWERING RESEARCH QUESTION

```
Call:
lm(formula = TemperatureC ~ Humidity + I(Humidity^2), data = weather.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9046 -1.6020  0.2689  1.1766  3.3584 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)      81.49      37.36   2.181   0.0342 *
Humidity       -168.31      88.91  -1.893   0.0645 .
I(Humidity^2)    90.63      52.66   1.721   0.0918 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.846 on 47 degrees of freedom
Multiple R-squared:  0.2808,	Adjusted R-squared:  0.2502 
F-statistic: 9.177 on 2 and 47 DF,  p-value: 0.000432
```

```
Call:
lm(formula = ApparentTemperatureC ~ Humidity + I(Humidity^2), 
    data = weather.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9768 -2.2010  0.3096  1.3638  7.0555 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      81.14      52.81   1.537    0.131
Humidity       -179.35     125.66  -1.427    0.160
I(Humidity^2)    98.82      74.43   1.328    0.191

Residual standard error: 2.609 on 47 degrees of freedom
Multiple R-squared:  0.1288,	Adjusted R-squared:  0.09175 
F-statistic: 3.475 on 2 and 47 DF,  p-value: 0.03913
```

The slope for each is given by the value to the right of the temperature variable. And the intercept above it, giving the following-

Equation for the Control

$$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 = (81.49) + (-168.31)(x) + (90.63)x^2$$

Equation for the Experiment

$$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 = (81.14) + (-179.35)(x) + (98.82)x^2$$

CONCLUSIONS

The model proved to be valid and satisfied the assumptions of the linear model.

The model does not serve as a very accurate predictor because of the wide ranging variance.

Furthermore humidity DOES affect the Apparent Temperature AS WELL as the real temperature.

Experiment could be expanded by looking into the effect of wind speed on temperature. It might also be improved by using Fahrenheit because it would give a larger range of values.