Rachael Soh
6 May 2020
Math189Z: Status report

Collecting weather data:

Since last time, I've found the WorldWeatherOnline historical weather data API wrapper "wwo_hist" to pull historical data of specific cities or regions. I'm using their 60 day free trial right now, which I think is fine when analyzing data now but is something we have to consider if we want to use this in our data visualizer. It gives a simple summary of weather factors we are interested in so it is extremely convenient to use in our current data analysis:

```
['date_time', 'maxtempC', 'mintempC', 'totalSnow_cm', 'sunHour',
'uvIndex', 'uvIndex.1', 'moon_illumination', 'moonrise', 'moonset',
'sunrise', 'sunset', 'DewPointC', 'FeelsLikeC', 'HeatIndexC',
'WindChillC', 'WindGustKmph', 'cloudcover', 'humidity', 'precipMM',
'pressure', 'tempC', 'visibility', 'winddirDegree', 'windspeedKmph']
```

CALIFORNIA

This time, I tried running a multivariate regression using the dependent variable R value which I calculated like so: (cases today - cases yesterday) / (cases yesterday - cases the day before). With the independent variables temperature (tempC), humidity, windspeedKmph and precipMM. I found that

Univariate Regression

Individual results:

- **tempC:**
  - slope: -0.01864194168757026
  - p-values: 0.34318292801644035
  - R^2: 0.01696776809983029
- **windspeed:**
  - p-values: 0.4403214225538251
  - R^2: 0.011277116428568741
  - Slope: 0.015525162620814367
- **humidity:**
  - p-values: 0.353740729263576
  - R^2: 0.01624643767517239
  - Slope: 0.00897228677665798
- **precipitation:**
  - p-values: 0.06428538008454321
  - R^2: 0.06311798707956556
  - Slope: 0.036715952819390954

When ran in multivariate regression:

In [264]:
```
model_lin = sm.OLS.from_formula("R ~ tempC + precipMM + windspeedKmph+humidity"
                                , data=CA_cases)
result_lin = model_lin.fit()
result_lin.summary()
```

Out[264]:

OLS Regression Results

Dep. Variable:	R	**R-squared:**	0.080
Model:	OLS	**Adj. R-squared:**	0.006
Method:	Least Squares	**F-statistic:**	1.080
Date:	Tue, 05 May 2020	**Prob (F-statistic):**	0.376
Time:	23:54:56	**Log-Likelihood:**	-73.070
No. Observations:	55	**AIC:**	156.1
Df Residuals:	50	**BIC:**	166.2
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.3905	1.282	1.084	0.283	-1.185	3.966
tempC	-0.0072	0.030	-0.244	0.809	-0.067	0.052
precipMM	0.0439	0.025	1.749	0.086	-0.006	0.094
windspeedKmph	0.0188	0.022	0.871	0.388	-0.025	0.062
humidity	-0.0089	0.016	-0.543	0.589	-0.042	0.024

Omnibus:	75.659	**Durbin-Watson:**	1.515
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	986.016
Skew:	3.662	**Prob(JB):**	7.75e-215
Kurtosis:	22.406	**Cond. No.**	596.

Finally, we see results that are consistent with what we have seen in news articles and other research done on the matter. It seems that increases in temperatures and humidity do decrease the number of cases. However, it is important to note that they are not significant! In fact, the covariate closest to being significant was rain. Another point to consider is that this data was analyzed for California, an extremely large land mass which gives us a worse estimate of weather, one thing I would like to do is analyze different cities in California, but I am still having trouble automating that process as the wwo_hist api does not have all cities or some are specific keys which I have to standardize.

Now, let's try it with interaction terms of each pair of weather factors tempC: humidity, tempC:precipMM, etc.

In [265]:
```
model_lin = sm.OLS.from_formula("R~tempC+precipMM+windspeedKmph+humidity+precipMM:windspeedKmph+precipMM:humidity+winc
result_lin = model_lin.fit()
result_lin.summary()
```

Out[265]:

OLS Regression Results

Dep. Variable:	R	**R-squared:**	0.484
Model:	OLS	**Adj. R-squared:**	0.366
Method:	Least Squares	**F-statistic:**	4.121
Date:	Wed, 06 May 2020	**Prob (F-statistic):**	0.000481
Time:	00:00:16	**Log-Likelihood:**	-57.175
No. Observations:	55	**AIC:**	136.4
Df Residuals:	44	**BIC:**	158.4
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.8846	3.676	1.329	0.191	-2.524	12.293
tempC	-0.0471	0.098	-0.479	0.635	-0.246	0.151
precipMM	1.7048	0.363	4.690	0.000	0.972	2.437
windspeedKmph	-0.4221	0.240	-1.759	0.086	-0.906	0.062
humidity	-0.0184	0.053	-0.349	0.729	-0.125	0.088
precipMM:windspeedKmph	-0.0037	0.004	-0.835	0.409	-0.013	0.005
precipMM:humidity	-0.0201	0.004	-4.831	0.000	-0.028	-0.012
windspeedKmph:humidity	0.0044	0.003	1.628	0.111	-0.001	0.010
tempC:humidity	-0.0024	0.002	-1.240	0.222	-0.006	0.001
tempC:windspeedKmph	0.0111	0.006	1.919	0.061	-0.001	0.023
tempC:precipMM	-0.0135	0.007	-1.862	0.069	-0.028	0.001

Omnibus:	16.206	**Durbin-Watson:**	1.832
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	24.055
Skew:	0.991	**Prob(JB):**	5.98e-06
Kurtosis:	5.563	**Cond. No.**	5.07e+04

Now, it seems that some of the interaction terms have a significant effect and some of our original covariates now are significant. We see that precipMM is now significant, as well as precipMM:humidity. Although precipMM increases the number of cases, precipMM:humidity decreases it. Although, it is worth considering that rain would result in a higher humidity so collinearity between covariates in the model should be accounted for. Although, I'm not quite sure how to do that here. Another factor to consider is that California's weather does not change very much and so it would make sense that it does not have a very large effect on the number of cases, which is why we will be analyzing Washington state next.

WASHINGTON
Univariate regression:
Individuals:
- tempC
p-values: 0.7465103771359007
R^2: 0.0018154399673399164
Slope: -0.012783595527740626
- windspeed:
p-values: 0.7860944202635958
R^2: 0.0012800032450079745
Slope: 0.011024402307496038
- humidity:
p-values: 0.8201350078973945
R^2: 0.0008986522517151205
Slope: -0.003415228186355623
- precipitation:
p-values: 0.8736916161177402
R^2: 0.0004393966234831305
Slope: -0.0034748390611039027

<u>Multivariate regression without interactions:</u>

Out[69]:

OLS Regression Results

Dep. Variable:	R	**R-squared:**	0.004
Model:	OLS	**Adj. R-squared:**	-0.069
Method:	Least Squares	**F-statistic:**	0.05003
Date:	Wed, 06 May 2020	**Prob (F-statistic):**	0.995
Time:	00:08:37	**Log-Likelihood:**	-106.01
No. Observations:	60	**AIC:**	222.0
Df Residuals:	55	**BIC:**	232.5
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.2413	1.638	0.758	0.452	-2.042	4.524
tempC	-0.0118	0.041	-0.285	0.777	-0.095	0.071
precipMM	-0.0042	0.029	-0.145	0.885	-0.062	0.054
windspeedKmph	0.0130	0.048	0.272	0.787	-0.083	0.109
humidity	0.0003	0.020	0.016	0.987	-0.040	0.041

Omnibus:	18.709	**Durbin-Watson:**	2.230
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	102.972
Skew:	0.356	**Prob(JB):**	4.36e-23
Kurtosis:	9.378	**Cond. No.**	589.

Once again, we see that none of our covariates are significant. Although an increase in temperature and precipitation seems to have decreased our R-values very slightly.

With interactions:

t[117]:

OLS Regression Results

Dep. Variable:	R	**R-squared:**	0.050
Model:	OLS	**Adj. R-squared:**	-0.181
Method:	Least Squares	**F-statistic:**	0.2169
Date:	Wed, 06 May 2020	**Prob (F-statistic):**	0.993
Time:	00:19:24	**Log-Likelihood:**	-92.178
No. Observations:	52	**AIC:**	206.4
Df Residuals:	41	**BIC:**	227.8
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0440	8.270	-0.005	0.996	-16.745	16.657
tempC	0.1507	0.467	0.323	0.749	-0.792	1.093
precipMM	0.0567	0.250	0.227	0.822	-0.448	0.561
windspeedKmph	0.0853	0.348	0.245	0.808	-0.617	0.788
humidity	-0.0366	0.114	-0.321	0.750	-0.267	0.194
precipMM:windspeedKmph	-0.0030	0.006	-0.496	0.623	-0.015	0.009
precipMM:humidity	-0.0006	0.003	-0.224	0.824	-0.006	0.005
windspeedKmph:humidity	0.0031	0.006	0.509	0.613	-0.009	0.015
tempC:humidity	0.0003	0.006	0.056	0.956	-0.011	0.012
tempC:windspeedKmph	-0.0137	0.015	-0.905	0.371	-0.044	0.017
tempC:precipMM	0.0021	0.009	0.228	0.821	-0.016	0.020

Omnibus:	18.075	**Durbin-Watson:**	2.202
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	87.891
Skew:	0.445	**Prob(JB):**	8.22e-20
Kurtosis:	9.306	**Cond. No.**	5.88e+04

This time, we see no significant covariates again, which was disappointing to see. Again, I think it is because the are is so large, it is difficult to get accurate weather data so I'd like to do it on multiple cities instead. Also, we see again a possibility of high collinearity in our data as the condition number is very high. Maybe we can look into non-linear models next. Interestingly, the R-value here is much lower than that in California's so perhaps weather may affect certain areas but not others.

MAIN TAKEAWAYS:
Overall, I think there are a few major problems with how I'm doing this analysis: 1. The area (state level) is much too large for accurate weather analysis and 2. There is a lot of collinearity between covariates: precipitation and humidity. So a few future steps I plan to take are:

1. Remove precipitation or humidity to remove collinearity. Unless there is a better way to handle these 2 terms. Further research must be done to find an appropriate measure.
2. This R value isn't the best to use for day to day comparisons as I often get very large spikes or infinite values because of either no increase or too much of an increase. Ex. Days 1,2 & 3 all have 1 case gives an infinite R value.
3. Automate the process of running the regression on city data instead of state data. Meaning, using the John Hopkins data, we can find the city in which those numbers were found. For example, in California we have SF, LA, SD, OC, etc. which would give much more accurate data. Then using these city names, get the weather data with our API. I'm not super sure how to do this without taking a long time and using a lot of space. I will probably write a python script that deletes the csv file right after it is read into the JupiterNotebook.
4. Include other explanatory variables like population density, GDP, number of hospitals per unit area, etc. to build a better predictive model.

General group goals:

- Further explore this relationship and perhaps use it for predictive purposes
- Discuss a good data visualization tool

Things I need advice on:

- For point 1, any advice on how to deal with collinearity and potential nonlinear models would be greatly appreciated!
- For point 3, is there any way to do it faster.
- Is there a way to get relationships for each city and then compare all the cities' relationships/data altogether? Or do I just combine all this information together and do one large regression on the larger dataset? (Instead of comparing 10 cities with 100 data points then combining their results, just do 1 collective 1000 point analysis)