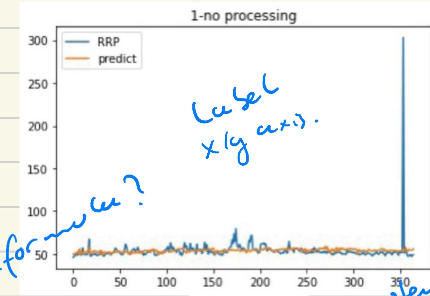
Question 1

Aisha Xu.

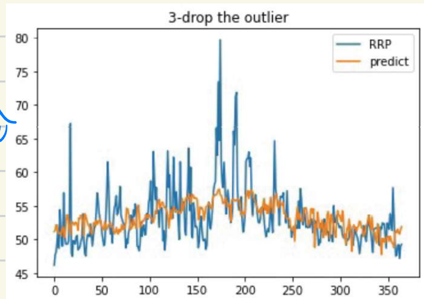
Since we only have RRP values for the whole 2013, we only took independent variables in SydTemp in 2013. From SydTemp file, we found columns ‘Month’,’Day’, ‘Maximum temperature’ and ‘Minimum temperature’ relevant in predicting RRP. Applying OLS directly, we plotted our predicted values and real values (see figure 1). This is model 1. Specify your model information?



label x/y axis.

Figure 1

→ systematic outlier detection way be better?



label

Figure 2

We observed that there was an outlier in December, we removed this outlier and fitted OLS again. The new model made the absolute value of log-likelihood, AIC and BIC decrease 33%, indicating a significant increase in the accuracy of the model without comprising simplicity. The new plot shows that the prediction can capture the ‘upward to July then downward’ trend slightly (see figure 2). This is model 2. → because of day or temperature?

We then convert the day to day-of-the-week, out of the interest that weekdays and weekends may differ. We plotted scattered graphs between each pair of variables.

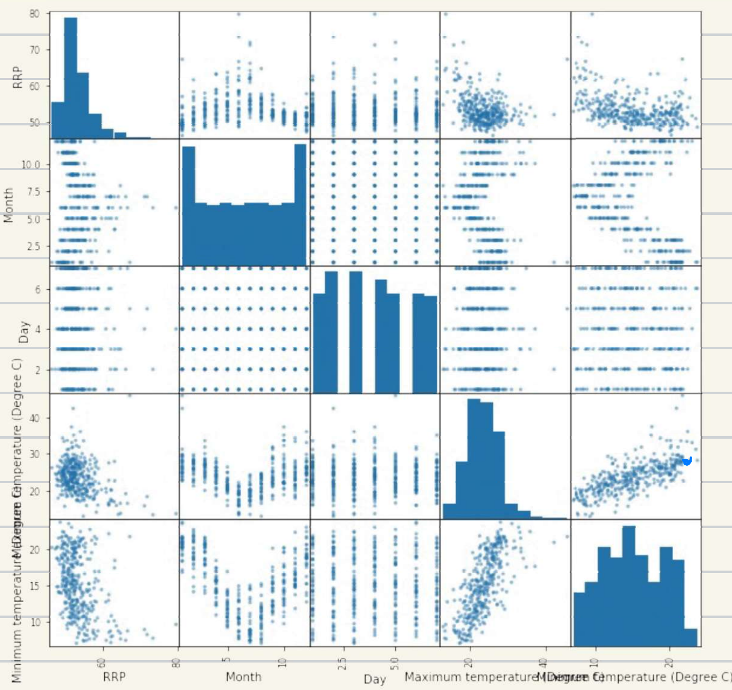


Figure 3

We observed no apparent change in RRP when day of the week changes. Our model didn’t get improved from this change - infinitesimal change in AIC, log-likelihood and BIC, but the probability that we tend to accept the coefficient of ‘day of the week’ is 0 rose to 0.485, while that of ‘day’ was nearly 0.

↳ use weekend as indicator (binary) variable

From figure 3, the scatter plot of RRP against Month indicated some quadratic relationship. We then tried to add ‘Month_sqr’ independent variable by squaring ‘Month’ values. There are slight improvement in AIC and log-Likelihood (around 1%), but the coefficient related to Month_sqr is negative (as illustrated by the scatter graph) and P > |t| is 0, so we decided to keep this column. Worth mentioning, the coefficient of max temperature and min temperature are both negative, scatter plots also illustrated negative correlations. This is model 3.

The P > |t| value for ‘maximum temperature’ is around 0.4 for previously tried models, we assumed that this column was not important and took it away, but there was no change to AIC so it didn’t help much. → so temperature does not have impact? Refer to main question.

→ good

Below are the ANOVA table, comparing (model 1, model 2), (model 2, model 3). We verified that there are decreases in ssr, so improvements in the accuracy of the model.

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	360.0	67803.200292	0.0	NaN	NaN	NaN
1	359.0	5188.744611	1.0	62614.45568	4332.18269	1.929972e-202

ANOVA (model 1, model 2)

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	360.0	67803.200292	0.0	NaN	NaN	NaN
1	359.0	5188.744611	1.0	62614.45568	4332.18269	1.929972e-202

ANOVA (model 2, model 3)

The coefficients of model 3 are:

	coef	std err	t	P>|t|	[0.025	0.975]
const	53.1480	2.304	23.069	0.000	48.617	57.679
x1	1.4647	0.418	3.503	0.001	0.643	2.287
x2	0.0649	0.022	2.926	0.004	0.021	0.109
x3	−0.0629	0.065	−0.974	0.331	−0.190	0.064
x4	−0.1514	0.094	−1.612	0.108	−0.336	0.033
x5	−0.1239	0.030	−4.084	0.000	−0.184	−0.064

Using this model, the predicted RRP values are:

```
array([49.74378474, 48.94450037, 49.85458569, 49.97397071, 50.02732142,
       49.89630329, 50.81268285])
```

Q2

In [21]:

```
def pielinear(x,k):
    import numpy as np

    n = len(x)
    m = len(k)
    M = np.zeros((n,m+2))
    M[:,0] = np.ones(n)
    M[:,1] = x

    X = np.array([x for i in range(m)])
    X = X.T
    K = np.array([k for j in range(n)])
    M[:,2:] = np.maximum(X-K,np.zeros((n,m)))
    return M
```

In [25]:

```python
import numpy as np
x = np.linspace(1,10,10)
k = np.linspace(-9,10,20)
M = pielinear(x,k)
M
```

```
array([[ 1.,  1., 10.,  9.,  8.,  7.,  6.,  5.,  4.,
3.,  2.,  1.,  0.,
         0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.],
       [ 1.,  2., 11., 10.,  9.,  8.,  7.,  6.,  5.,
4.,  3.,  2.,  1.,
         0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.],
       [ 1.,  3., 12., 11., 10.,  9.,  8.,  7.,  6.,
5.,  4.,  3.,  2.,
         1.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.],
       [ 1.,  4., 13., 12., 11., 10.,  9.,  8.,  7.,
6.,  5.,  4.,  3.,
         2.,  1.,  0.,  0.,  0.,  0.,  0.,  0.,  0.],
       [ 1.,  5., 14., 13., 12., 11., 10.,  9.,  8.,
7.,  6.,  5.,  4.,
         3.,  2.,  1.,  0.,  0.,  0.,  0.,  0.,  0.],
       [ 1.,  6., 15., 14., 13., 12., 11., 10.,  9.,
8.,  7.,  6.,  5.,
         4.,  3.,  2.,  1.,  0.,  0.,  0.,  0.,  0.],
       [ 1.,  7., 16., 15., 14., 13., 12., 11., 10.,
9.,  8.,  7.,  6.,
         5.,  4.,  3.,  2.,  1.,  0.,  0.,  0.,  0.],
       [ 1.,  8., 17., 16., 15., 14., 13., 12., 11., 1
0.,  9.,  8.,  7.,
         6.,  5.,  4.,  3.,  2.,  1.,  0.,  0.,  0.],
       [ 1.,  9., 18., 17., 16., 15., 14., 13., 12., 1
1., 10.,  9.,  8.,
         7.,  6.,  5.,  4.,  3.,  2.,  1.,  0.,  0.],
       [ 1., 10., 19., 18., 17., 16., 15., 14., 13., 1
2., 11., 10.,  9.,
         8.,  7.,  6.,  5.,  4.,  3.,  2.,  1.,  0.]])
```

In [38]:

```
x2 = np.random.rand(5)
print('x2 =',x2)
k2 = np.linspace(0,1,3)
print('k2 =', k2)
M2 = pielinear(x2,k2)
M2
```

```
x2 = [0.57067815 0.94132958 0.12579889 0.46471408 0.9960
6173]
k2 = [0.  0.5 1. ]
```

```
array([[1.        , 0.57067815, 0.57067815, 0.07067815,
0.        ],
       [1.        , 0.94132958, 0.94132958, 0.44132958,
0.        ],
       [1.        , 0.12579889, 0.12579889, 0.        ,
0.        ],
       [1.        , 0.46471408, 0.46471408, 0.        ,
0.        ],
       [1.        , 0.99606173, 0.99606173, 0.49606173,
0.        ]])
```

Q3 MSE of θ and σ^2 can be approximated by

$$\mathbb{E}\left[(\theta-\sigma^2)^2\right]$$
$$=\mathbb{E}\left[\left[\alpha\sum_i (X-\bar{X})^2-\sigma^2\right]^2\right]$$
$$=\mathbb{E}\left[\alpha^2\left[\sum_i (X-\bar{X})^2\right]^2-2\sigma^2\alpha\sum_i (X-\bar{X})^2+\sigma^4\right]$$
$$=\alpha^2\,\mathbb{E}\left[\sum_i (X-\bar{X})^2\right]^2-2\sigma^2\alpha\,\mathbb{E}\left(\sum_i (X-\bar{X})^2\right)+\sigma^4$$

$$Var\left[\frac{1}{n}\sum_i (X-\bar{X})^2\right]=\frac{1}{n^2}Var\left(\sum_i (X-\bar{X})^2\right)$$
$$Var\left[\sum_i (X-\bar{X})^2\right]=\mathbb{E}\left[\sum_i (X-\bar{X})^2\right]^2-\left[\mathbb{E}\left(\sum_i (X-\bar{X})^2\right)\right]^2$$

Since $\mathbb{E}\left(\frac{\sum_i (X_i-\bar{X})^2}{n-1}\right)=\sigma^2$. $\Rightarrow \mathbb{E}\left(\sum_i (X_i-\bar{X})^2\right)=(n-1)\sigma^2$.

Moreover, $Var\left[\sum_i (X-\bar{X})^2\right]=(n-1)^2\left(\frac{\gamma}{n}+\frac{2}{n-1}\right)$

$$\Rightarrow (n-1)^2\left(\frac{\gamma}{n}+\frac{2}{n-1}\right)=\mathbb{E}\left(\sum_i (X-\bar{X})^2\right)^2-(n-1)^2\sigma^2$$
$$\mathbb{E}\left(\sum_i (X-\bar{X})^2\right)^2=\frac{\gamma(n-1)^2}{n}+2(n-1)+(n-1)^2\sigma^2$$

$$e:=\mathbb{E}\left((\theta-\sigma^2)^2\right)=\alpha^2\left[\frac{\gamma(n-1)^2}{n}+2(n-1)+(n-1)^2\sigma^4\right]-2\sigma^2\alpha(n-1)\sigma^2+\sigma^4$$

(2) $$\frac{\partial e}{\partial\alpha}=2\alpha\left[\frac{\gamma(n-1)^2}{n}+2(n-1)+n^2\sigma^4\right]-2\sigma^4(n-1)=0$$

$$\alpha=\frac{\sigma^4(n-1)}{\frac{\gamma(n-1)^2}{n}+2(n-1)+(n-1)^2\sigma^4}=\frac{\sigma^4 n}{\gamma(n-1)+2n+n(n-1)\sigma^4}$$

$$e_{min}=\alpha\sigma^4(n-1)-2\sigma^2\alpha(n-1)\sigma^2+\sigma^4=-\alpha\sigma^4(n-1)+\sigma^4=\sigma^4(1-\alpha n+\alpha)$$

$$X \sim N(\mu, \sigma^2)$$

$$\frac{X-\mu}{\sigma} \sim N(0,1). \qquad \frac{(X-\mu)^2}{\sigma^2} \sim \chi_1^2,$$

$$\Rightarrow E\left(\frac{(X-\mu)^2}{\sigma^2}\right) = 1, \quad Var\left(\frac{(X-\mu)^2}{\sigma^2}\right) = 2.$$

$$E\left(\frac{X-\mu}{\sigma^2}\right)^4 - \left(E\left(\frac{(X-\mu)^2}{\sigma^2}\right)\right)^2 = 2$$

$$\Rightarrow E\left(\frac{X-\mu}{\sigma^2}\right)^4 = 3 \quad \Rightarrow E(X-\mu)^4 = 3\sigma^8$$

$$\gamma = \frac{3\sigma^8}{\sigma^4} - 3 = 3\sigma^4 - 3$$

$$\alpha = \frac{\sigma^4 n}{(3\sigma^4-3)(n-1) + 2n + n(n-1)\sigma^4}$$

$$E(\theta) = \alpha E\left[\sum_i (X_i - \bar{X})^2\right] = \alpha(n-1)\sigma^2 = \frac{\sigma^4 n(n-1)}{(3\sigma^4-3)(n-1)+2n+n(n-1)\sigma^4}\sigma^2$$

$$|E(\theta) - \sigma^2| = \left|\frac{\sigma^4 n(n-1) - (3\sigma^4-3)(n-1) - 2n - n(n-1)\sigma^4}{(3\sigma^4-3)(n-1)+2n+n(n-1)\sigma^4}\sigma^2\right|$$

$$= \frac{3(\sigma^4-1)(n-1)+2n}{3(\sigma^4-1)(n-1)+2n+n(n-1)\sigma^4}\sigma^2. = \frac{1}{1+\frac{n(n-1)\sigma^4}{3(n-1)(\sigma^4-1)+2n}}\sigma^2.$$

$$E((\theta-\sigma^2)^2) - \left(\frac{\gamma}{n} + \frac{2}{n-1}\right) = \sigma^4(1 - \alpha n + \alpha) - \frac{\gamma}{n} - \frac{2}{n-1}$$

$$= \sigma^4 - \alpha(n-1)\sigma^4 - \frac{\gamma}{n} - \frac{2}{n-1}$$

$$= \sigma^4 - \frac{\sigma^4 n(n-1)}{(3\sigma^4-3)(n-1)+2n+n(n-1)\sigma^4}\sigma^4 - \frac{3(\sigma^4-1)}{n} - \frac{2}{n-1}.$$