Linear Regression model to predict blooming of Cherry Blossom

The goal of this project is to build the best possible linear model to explain the day of peak bloom. Using the dataset "CherryData 2016" data from 1921 to 2015 (note do not include 2016 data) build a multiple linear regression model to predict the Peak Bloom Date.***

--The given dataset contains the data of peak date of bloom, temperature and snow conditions of the month January February and March from 1921 to 2016.One of the attribute of the dataset namely "***Day.Peak.Bloom***" refers to number of days elapsed since 1st January of a given year and date of peak bloom. The aim of the project is to find how independent variables (IV) like temperature and snow conditions of the month January February and March effects the dependent variable (DV) "***Date of peak bloom.***" In order to do so we establish a relationship among **Independent variables** like "*January temperature", "January snow", "February temperature", "February Snow", "March temperature" "March snow"* and **Dependent Variable** (which in this case is *Day peak bloom*), since it directly gives the difference between 1st January of a given year and the date of peak bloom. Hence the exact date of peak bloom can be derived from there.

To find the relationship of how these Independent variables affects the dependent variable we do a multiple regression analysis in R. To do so firstly a subset from the actual dataset is created where in data from only 1921-2015 is present. Multiple regression in R is done by using the following command in R

```
summary(lm(a$Day.Peak.Bloom ~ a$JanTemp+a$FebTemp+a$MarTemp+a$JanSnow+a$FebSn
ow+a$MarSnow,data=a))
```

```
Call:
lm(formula = a$Day.Peak.Bloom ~ a$JanTemp + a$FebTemp + a$MarTemp +
    a$JanSnow + a$FebSnow + a$MarSnow, data = a)

Residuals:
     Min       1Q   Median       3Q      Max
-10.8356  -2.2316   0.0441   1.9986  13.1526

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 184.889154   8.517148  21.708  < 2e-16 ***
```

```
a$JanTemp    -0.046435   0.105489  -0.440    0.661    
a$FebTemp    -0.615685   0.131919  -4.667 1.09e-05 ***
a$MarTemp    -1.419101   0.124447 -11.403  < 2e-16 ***
a$JanSnow    -0.039801   0.073411  -0.542    0.589    
a$FebSnow     0.006768   0.071139   0.095    0.924    
a$MarSnow    -0.019790   0.136244  -0.145    0.885    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.843 on 88 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.7352,   Adjusted R-squared:  0.7171 
F-statistic: 40.71 on 6 and 88 DF,  p-value: < 2.2e-16
```

Here all the Independent Variables are utilized to get the full model equation. Multiple linear regression is a generalization of simple linear regression to handle multiple IVs. We define a linear model:

Y = B0 + B1X1 + B2X2 + B3X3 + ::: + error (where Y is the dependent variable and X1,X2,X3… are independent variables and B0,B1,B2…. are estimated intercepts of the line)

Hence the full model appears as

$\hat{y}$ =184.889154-0.046435(JanTemp)-0.615685(FebTemp)-1.419101(MarTemp)-0.039801(JanSnow) + 0.006768(FebSnow) -0.019790(MarSnow)

In this model $\hat{y} = Day\ Peak\ Bloom$ from which date of peak bloom can be derived. Analyzing the result from multiple regression we realize that not all IV are significantly related to our dependent variable.

The overall hypothesis test results are given by F(688) = 40.71, p-value: < 2.2e-16. This hypothesis test evaluates the full model in comparison to the baseline intercept only model. The null hypothesis is that both models explain equal variance in the Dependent Variable. The alternative hypothesis is that our model with the additional predictors explains more variance than the baseline model. In this case p ≤ α so we reject the null hypothesis and conclude that our model explains more variance. In other words, the results of this hypothesis test shows that the model has some predictive ability.

Once we have established that the overall model is significant, we can examine each predictor individually to see if it predicts the Dependent Variable. We evaluate the hypothesis test for each coefficient B. The p value of ***FebTemp*** (*Temperature of Feb*)and ***MarTemp*** ***(****Temperatue of March****)*** are less than α=0.05. ***Hence we can conclude that FebTemp*** and ***MarTemp are related to the dependent***

variable. Whereas the remaining Independent Variables namely JanTemp** (Temperature in January)**, JanSnow** (Snow in January)**, FebSnow** (Snow in February) **and MarSnow** (Snow in March) **are not related to the dependent variable because they have p greater than or comparable to that of 0.05.

Finally, we rerun the model with only significant variables (FebTemp and MarTemp) and we use the following R command.

```
summary(lm(a$Day.Peak.Bloom ~ a$FebTemp+a$MarTemp,data=a))

Call:
lm(formula = a$Day.Peak.Bloom ~ a$FebTemp + a$MarTemp, data = a)

Residuals:
Min       1Q   Median       3Q      Max
-10.4823  -2.2295   0.1214   1.7676  13.2150

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 182.85436    5.63635  32.442  < 2e-16 ***
a$FebTemp    -0.63062    0.09705  -6.498 4.12e-09 ***
a$MarTemp    -1.40385    0.10455 -13.428  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.766 on 92 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared:  0.7341,   Adjusted R-squared:  0.7283
F-statistic:   127 on 2 and 92 DF,  p-value: < 2.2e-16
```

The interpretation of the result is ***FebTemp*** (*Temperature of February*) is negatively related to day peak bloom . With each degree increase in February temperature the bloom date comes closer to January 1st of the year by 0.63062 points. ***MarTemp*** (*Temperature of March*) is also negatively related to Day peak Bloom. With each degree increase in March temperature the bloom date comes closer to January 1st of the year by 1.40385 points.

Hence now the new model is

$\hat{y}$ **=182.854 -0.63062(*FebTemp*) -1.40385(*MarTemp*)**

In order to understand how much **variance** is explained by the model, we compare the ***Multiple R squared value and adjusted R value of full model and new model***. Multiple R-squared is used for evaluating how well our model fits the data. It tells how much of the variance in the dependent variable (the predicted variable) can be explained by the independent variables (the predictor variables). For example, an R-squared value of 0.75 implies that the model can explain three-quarters of the variation in the outcome. Every time an independent variable is added to the model, the R-squared value will increase. Adjusted R-squared also provides the same information as R-squared but adjusts for the number of terms in the model. It does not monotonically increase like R-squared but increases only when the new variable actually has an effect on the predicted value. It decreases when the new variable does not have any real impact on the predicted value.

Comparing the full model Multiple R squared value with the new model; **Full model Multiple R-squared: 0.7352 and New Model Multiple R-squared: 0.7341.** Both of which are comparable to one another.

Full Model Adjusted R-squared: 0.7171 and New Model Adjusted R-squared: 0.7283. Comparing both we realize that the new model has a better adjusted R squared value which is greater than full model; hence the new model fits the data better than the full model.

The new **model** $\hat{y}$ **=182.854 -0.63062(*FebTemp*) -1.40385(*MarTemp*)** will give better predictions than the full model equations.

Assumptions:

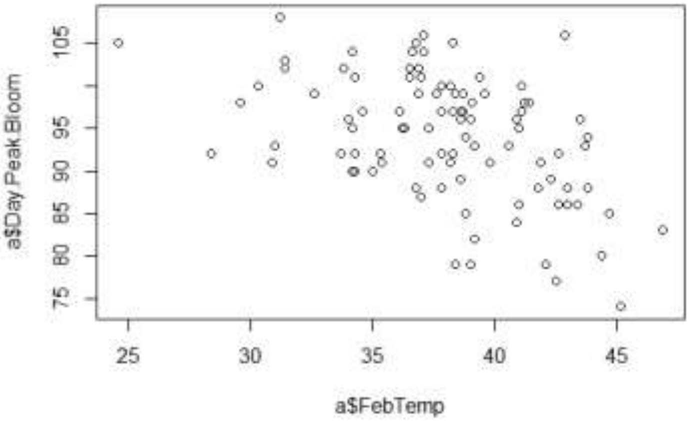
Multicollinearity looks at correlations between variables. Finding collinearity between the independent variables from our new model, namely February Temperature and March Temperature we use the following R command

```
cor.test(a$MarTemp,a$FebTemp)

        Pearson's product-moment correlation

data:  a$MarTemp and a$FebTemp
t = 1.4475, df = 93, p-value = 0.1511
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.05474442  0.33981338
sample estimates:
      cor
0.1484363
```

A collinearity coefficient of 0.1484 shows that both these variables are not very closely correlated.

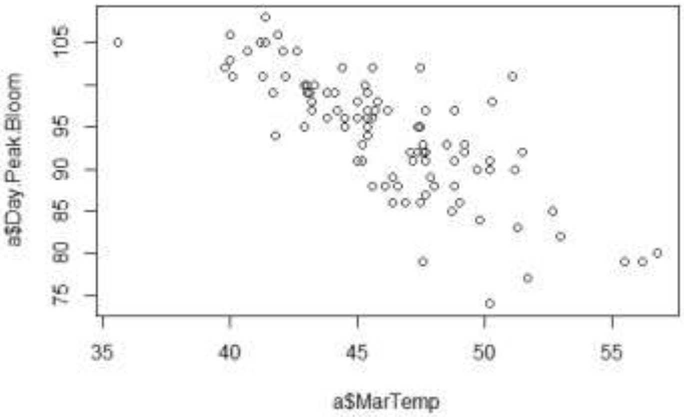
To check for Non linearity Plot each X against your Y using a scatterplot to check for non-linearity.

Plotting February Temperature against day peak bloom



This graph is scattered although majority of data points are aligning to a negative slope still its quite spread out, hence its nonlinear.

Plotting March Temperature against day peak bloom we get the following graph.



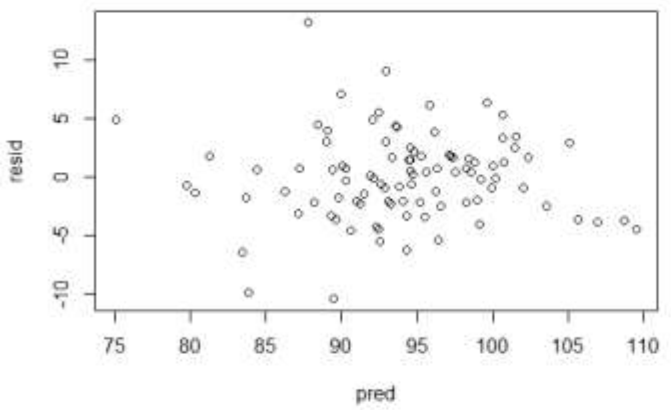
This graph shows linearity with a negative slope.

Looking for Outliers find out observations that are implausible by graphing, examining the x, y values, and by computing diagnostic statistics. Check outliers by getting observations that are 3 standard deviations away from the predicted values. the following command is used in R

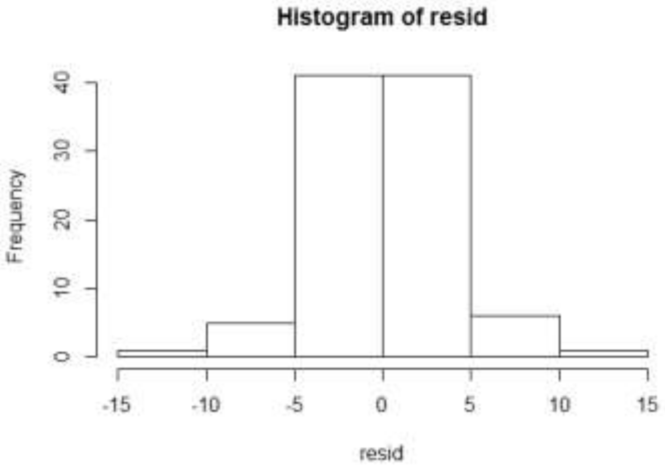
```
pred=m$fitted.values
> resid=m$residuals
> resid.sd=sd(resid)
> resid[abs(resid)>=3*resid.sd]
      53
13.21504
```

This shows only one value is an outlier and hence can be ignored. 3 times sd is 13.21504

Constant Error Variance in error should be about the same at all levels of X. Plotting predicted values as (X) and r residuals as (Y) using a scatterplot we check whether our residuals vary consistently at all predicted values. If they do not vary evenly, then this is a violation of constant error. However looking at the graph it shows that it varies.



Normality of Errors -Errors should be normally distributed. Create a histogram of residuals to check for non-normality of errors. You can also test normality using the Shapiro Wilk hypothesis tests.



Looking at the histogram it shows us normal distribution of the residual.

Following R command is used for shapiro wilk normality test

```
shapiro.test(resid)

        Shapiro-Wilk normality test

data:  resid
W = 0.98127, p-value = 0.1921
```

Independence of Errors -Errors at one level of X should not be related to errors at another level of X

```
m=lm(a$Day.Peak.Bloom~a$FebTemp+a$MarTemp,data=a)

> durbinWatsonTest(m)

 lag Autocorrelation D-W Statistic p-value

   1      0.2203173      1.557957   0.022

 Alternative hypothesis: rho != 0.
```

In R, the function durbinWatson Test() from car package tells if the residuals from a linear model are correlated or not

- The null hypothesis is that there is no correlation among residuals they are independent
- The alternative hypothesis is that residuals are correlated.

As p value is near 0 it means one can reject null.

Here P is 0.022 less than 0.05 .Hence we reject the null i.e. independent variables are correlated .

To predict the Peak Bloom Date for 2016

Using new model equation $\hat{y}$ =**182.854 -0.63062(*FebTemp*) -1.40385(*MarTemp*) ;** putting in values for February temperature = 39.9 and march temperature = 53.5 we derive the day peak bloom as follows

$\hat{y}$ =182.854 -0.63062(39.9) -1.40385(53.5)

$\hat{y}$ =182.854 – 25.1617 –75.105

$$\boldsymbol{\hat{y} = 82.58}$$

Day peak bloom as mentioned earlier is the difference of days between jan 1st and the peak bloom date. Hence adding 82.58~ 83 days to 1st Jan 2016 we get ***a date of 24th March 2016 as the date with peak bloom.***

Comparing this date with the data of peak bloom date wherein the actual dataset containing 2016 data is present, *25th march 2016 is date of peak bloom whereas according to our revised regression model 24th March 2016 is the date of peak bloom, which is comparable to the original.*