Analysis of climate-crop yield relationships in Canada with Distance Correlation

by

Yifan Dai

(Under the Direction of Lynne Seymour)

Abstract

Distance correlation is a new measure of relationships between random variables introduced by Szekely et al (2007). Distance correlation is determined by the distances over all pairs of points while Pearson correlation is determined by the distance between each point and mean. Therefore, distance correlation has properties of a true dependence measure. In this thesis, we build 6 best models for 6 types of crop (barley, canola, flax, oats, pea and spring wheat) in Regina, Saskatchewan by using distance correlation. Despite the complexity of other factors, we show how temperature and precipitation affect crop yield in the Canadian growing season from April to September. Equipped with this information, we are able to estimate the future viability as well as the supply of crops in Canada in response to climate change.

Key Words: Dependence, Distance correlation, Climate Change, Model Selection

Analysis of climate-crop yield relationships in Canada with Distance Correlation

by

Yifan Dai

A Thesis Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

MASTER OF SCIENCE

Athens, Georgia

2015

Analysis of climate-crop yield relationships in Canada with Distance Correlation

by

Yifan Dai

Approved:

Major Professor: Lynne Seymour

Committee: Cheolwoo Park

Liang Liu

Electronic Version Approved:

Suzanne Barbour

Dean of the Graduate School

The University of Georgia

December 2015

ACKNOWLEDGMENTS

First, I would like to thank Budda. You warm my heart and show me the way and provide me with strength to overcome difficulties. With your blessings, I achieve tremendous heights in both spiritual and academic.

Second, I would like to thank my parents. Thank you for your unconditional love and supports. Your teachings make me strong.

Third, I would like to thank Dr. Seymour as my major professor, for her patience, motivation and immense knowledge. I could not have imagined how difficult would it be without her directions and continuous supports. I would also like to thank Dr. Park and Dr. Liu for being my committee members.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

In the chapter 2, we introduce the definition of distance correlation as well as some properties of it. Then we examine distance correlation R and Pearson correlation r for both linear and nonlinear relationships by simulating data. In nonlinear cases, we investigate six patterns (wave, circle, quadratic, diamond, X-shape and cluster) and compare several statistics such as Pearson correlation r and distance correlation R, Pearson covariance cov and distance covariance $d\,\mathrm{cov}$, p-value for both Pearson correlation test ($H_0 : \rho = 0;\ H_a : \rho \neq 0$)and distance correlation test ($H_0 : f_{X,Y} = f_X f_Y;\ H_a : f_{X,Y} \neq f_X f_Y$).

In the chapter 3, we perform exploratory data analysis. The crop yield dataset includes yield information on 15 food crops observed in different regions within the provinces of Alberta, Saskatchewan and Manitoba, accounting for 84% of the total crop land in Canada. In this thesis, we investigate 6 different types of crops (barley, canola, flax, oats, peas and spring wheat) in Regina of Saskatchewan. The climate variables were recorded daily and include maximum temperature, minimum temperature, mean temperature, rainfall, snowfall and total precipitation (in which snowfall is converted to a rain equivalent). The daily climate record in Regina spans over 100 years, which is from 1891 to 2008. The growing and harvest season in Canada is from April to September while the fallow is from October to March of the next year. Despite the

complexity of other factors, we show that how temperature and precipitation affect crop yield in growing season from April to September.

In chapter 4, we use both distance correlation and Pearson correlation to help us select significant variables. We analyze 6 different crops which are barley, canola, flax, pea, oats and spring wheat. For each crop, we start with two full models. The first full model (Full 1) includes significant variables only. The second full model (Full 2) includes not only significant variables but also their corresponding interaction terms. Further, we also consider cumulative precipitation in February and March (fallow months) and include this term in both Full 1 and Full 2. Then we perform backward selection, forward selection, stepwise selection from null, and stepwise selection from full on both of them. We take consideration of various factors such as RMSE, RSS, AIC, BIC, R square, number of parameters and so on and select the best model from each starting model. Then we compare the two best models to get the final best model for each crop. Our cutoff line for p -value is 0.1which is α level.

In chapter 5, we make a conclusion and summarize all the best models.

In the last chapter, we give the further directions that how to make our models more accurate.

CHAPTER 2

DISTANCE CORRELATION

2.1 INTRODUCTION

Distance correlation is a new concept which was first introduced by Szekely, Rizzo and Bakirov (2007). It is a new measurement of relationships between random vectors. Distance correlation is determined by the distances over all pairs of points while Pearson correlation is determined by the distances between each point and mean. Therefore, distance correlation has properties of a true dependence measure, analogous to Pearson correlation r . For all distributions with finite first moments, distance correlation $R(X,Y)$ is defined for X and Y in arbitrary dimensions. Distance correlation R satisfies $0 \le R \le 1$, and we have $R = 0$ if and only if X and Y are independent while $R = 1$ if and only if X and Y are perfectly dependent. However, since distance correlation R is non-negative number, we are unable to tell whether two vectors are positively or negatively correlated to each other. Distance correlation can only tell us how strongly two vectors are correlated to each other.

2.2 DEFINITION

2.2.1 Distance dependence statistics

For an observed random sample $(X,Y)=\{(X_k,Y_k):k=1,...,n\}$ from the joint distribution of random vectors X in R^p and Y in R^q, define

$$a_{kl}=\left|X_k-X_l\right|_p, \qquad \bar{a}_{k\cdot}=\frac{1}{n}\sum_{l=1}^{n}a_{kl}, \qquad \bar{a}_{\cdot l}=\frac{1}{n}\sum_{k=1}^{n}a_{kl},$$

$$\bar{a}_{\cdot\cdot}=\frac{1}{n^2}\sum_{k,l=1}^{n}a_{kl}, \qquad A_{kl}=a_{kl}-\bar{a}_{k\cdot}-\bar{a}_{\cdot l}+\bar{a}_{\cdot\cdot},$$

$$b_{kl}=\left|Y_k-Y_l\right|_q, \qquad \bar{b}_{k\cdot}=\frac{1}{n}\sum_{l=1}^{n}b_{kl}, \qquad \bar{b}_{\cdot l}=\frac{1}{n}\sum_{k=1}^{n}b_{kl},$$

$$\bar{b}_{\cdot\cdot}=\frac{1}{n^2}\sum_{k,l=1}^{n}b_{kl}, \qquad B_{kl}=b_{kl}-\bar{b}_{k\cdot}-\bar{b}_{\cdot l}+\bar{b}_{\cdot\cdot},$$

where $k,l=1,...,n.$

2.2.2 Distance covariance

The empirical distance covariance $V_n(X,Y)$ is the nonnegative number defined by

$$V_n^2(X,Y)=\frac{1}{n^2}\sum_{k,l=1}^{n}A_{kl}B_{kl}.$$

Similarly, distance variance $V_n(X)$ is the nonnegative number defined by

$$V_n^2(X)=V_n^2(X,X)=\frac{1}{n^2}\sum_{k,l=1}^{n}A_{kl}^2.$$

$$V_n^2(Y)=V_n^2(Y,Y)=\frac{1}{n^2}\sum_{k,l=1}^{n}B_{kl}^2.$$

2.2.3 Distance correlation

The empirical distance correlation $R_n(X,Y)$ is the square root of

$$R_n^2(X,Y) = \begin{cases} \dfrac{V_n^2(X,Y)}{\sqrt{V_n^2(X)V_n^2(Y)}}, & V_n^2(X)V_n^2(Y) > 0. \\ 0, & V_n^2(X)V_n^2(Y) = 0. \end{cases}$$

We may find R_n is easy to compute and it is a good empirical measure of dependence. It is defined by the distances over all pairs of points instead of the distances between points and mean.

Note that the statistic $V_n(X) = 0$ if and only if every sample observation is identical. Indeed, if $V_n(X) = 0$, then $A_{kl} = 0$ for $k,l = 1,...,n.$ Thus $0 = A_{kk} = -\bar{a}_{k\cdot} - \bar{a}_{\cdot k} + \bar{a}_{\cdot\cdot}$ implies that $\bar{a}_{k\cdot} = \bar{a}_{\cdot k} = \bar{a}_{\cdot\cdot}/2$, and $0 = A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot} = a_{kl} = \left|X_k - X_l\right|_p$, so we have $X_1 = \cdots = X_n$.

2.3 SIMULATION AND COMPARISON

Pearson Correlation r, unfortunately, is not a useful measure of dependency in general. A lack of Pearson correlation or even $r = 0$ does not mean there is no relationship between two variables. Pearson correlation r works well as a measure of linear dependency but not for nonlinear relationships. Distance correlation R, as we mentioned above, is computed by the distances over all pairs of points instead of the distances between points and the mean. Therefore, distance correlation has properties of a true dependence measure for both linear and nonlinear relationships.

In this section, we examine distance correlation R and Pearson correlation r for both linear and nonlinear relationships by simulating data. In nonlinear cases, we investigate six patterns

(wave, circle, quadratic, diamond, X-shape and cluster) and compare several statistics such as Pearson correlation r and distance correlation R, Pearson covariance cov and distance covariance $d\,\mathrm{cov}$, p-value for both Pearson correlation test ($H_0 : \rho = 0;\ H_a : \rho \neq 0$)and distance correlation test ($H_0 : f_{X,Y} = f_X f_Y;\ H_a : f_{X,Y} \neq f_X f_Y$).

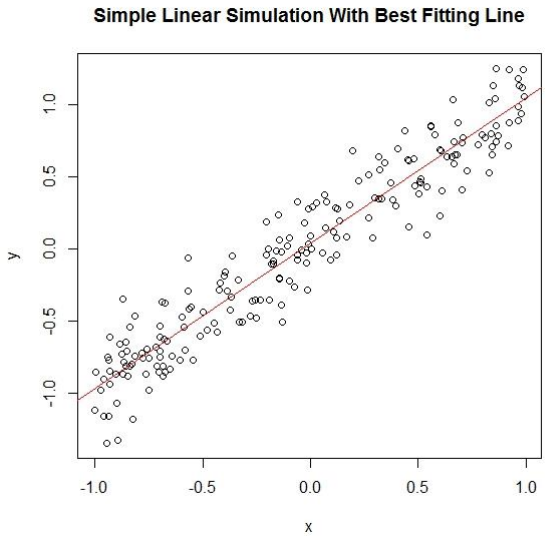
2.3.1 Linear relationship

We propose the following simple linear model:

$$y = x + \varepsilon$$

where x is sampled from uniform distribution with minimum = -1 and maximum = 1. The error term ε is from a normal distribution with mean = 0 and standard deviation = 0.2.

We use R to randomly sample 200 times and calculate Pearson correlation, distance correlation, Pearson covariance, distance covariance, p-value of Pearson correlation test and p-value of distance correlation, respectively. Then, we repeat this procedure for 100 times and we calculate the mean, standard deviation and quantiles for each statistic. Figure 2.3.1 is a sample scatter plot with best fitting line.



Below is table 2.3.1 which is the summary of statistics.

Table 2.3.1 Summary of Statistics for linear relationships

		Mean	Standard Dev.	Quant. at 2.5%	Quant. at 97.5%
Correlation	Pearson	0.9442	0.0071	0.9294	0.9579
	Distance	0.9387	0.0088	0.9206	0.9555
Covariance	Pearson	0.3302	0.0223	0.2878	0.3764
	Distance	0.3968	0.0206	0.3584	0.4382
p -value	Pearson	<0.0001	<0.0001	<0.0001	<0.0001
	Distance	<0.0001	<0.0001	<0.0001	<0.0001

In contrast, both Pearson correlation and distance correlation detect the true dependence between simulated data *x* and *y* very well. The mean of Pearson correlation and the mean of distance correlation are close to each other and both of them are greater than 0.9 which indicate that *x* and *y* are strongly correlated. Mean *p* -values for both Pearson correlation test and distance correlation test are extremely close to 0. In this case, we have sufficient evidence to reject the null hypothesis (H_0 : *x* and *y* are uncorrelated.) and conclude that *x* and *y* are related to each other (H_a : *x* and *y* are correlated.). Therefore, if two vectors are linearly related, we may make a similar conclusion from either Pearson correlation or distance correlation.

2.3.2 Wave Shape

We propose the following nonlinear model:

$$y = 2\sin(\frac{1}{2}x + \pi) + \varepsilon$$

where x is sampled from uniform distribution with minimum = - π and maximum = 7π.The error term ε is sampled from a normal distribution with mean = 0 and standard deviation = 1 (more noise).

We sample and compute statistics as before. Figure 2.3.2 is a sample scatter plot with best fitting line.

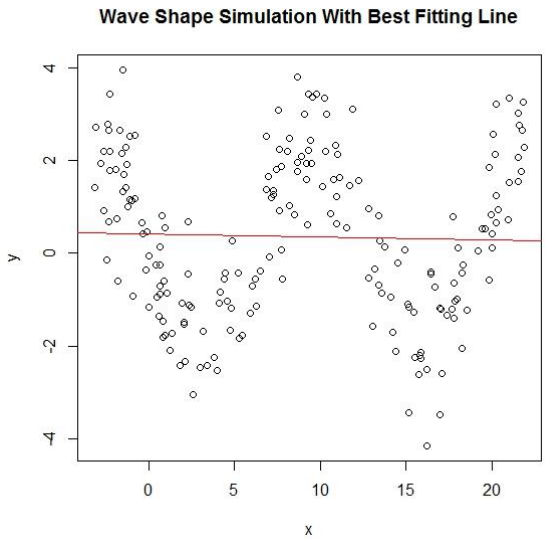


Figure 2.3.2 Wave Shape Simulation

Below is table 2.3.2 which is the summary of statistics.

Table 2.3.2 Summary of Statistics for Wave relationships

		Mean(ε =0)	Standard Dev.	Quant. at 2.5%	Quant. at 97.5%
Correlation	Pearson	-0.0036(0.007)	0.0668	-0.1319	0.1295
	Distance	0.2309(0.270)	0.0153	0.2095	0.2632
Covariance	Pearson	-0.0419	0.8412	-1.6882	1.6268
	Distance	0.5758	0.0435	0.5066	0.6660
p -value	Pearson	0.5113	0.2667	0.0499	0.9179
	Distance	0.0073	0.0037	0.0050	0.0150

Pearson correlation r is extremely close to 0 which tells us that there is no significant relationship between x and y. In contrast, distance correlation R does tell us that x and y might be related to each other although distance correlation R is not big. Further, the mean p-value for the Pearson correlation test is 0.5113. We have insufficient evidence to reject the null hypothesis (H_0 : x and y are uncorrelated.). In contrast, the mean p-value for the distance correlation test is 0.0073. At α level 0.1, we have sufficient evidence to reject the null hypothesis (H_0 : x and y are independent.). We have totally different conclusions when x and y are wave (nonlinear) correlated. Therefore, we conclude that distance correlation R outperforms Pearson correlation r at least if two vectors are nonlinearly correlated. Distance correlation may be able to detect the true nonlinearly dependence between x and y.

2.3.3 Circle Shape

We propose the following nonlinear model:

$$x^2 + y^2 = 1$$

where $x = \cos\theta + \varepsilon_1$ and $y = \sin\theta + \varepsilon_2$, θ is sampled from uniform distribution with minimum = 0 and maximum = 2π. The random error terms ε_1 and ε_2 are sampled from normal distribution with mean = 0 and standard deviation = 0.1 (less noise).

We sample and compute the same statistics as before. Figure 2.3.3 is a sample scatter plot with best fitting line.

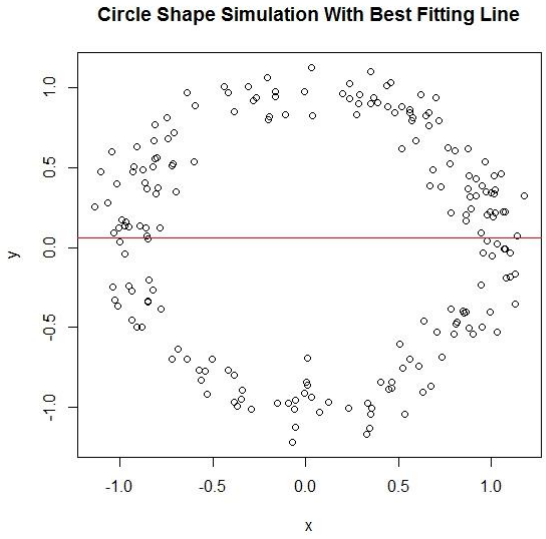


Figure 2.3.3 Circle Shape Simulation

Below is table 2.3.3 which is the summary of statistics.

Table 2.3.3 Summary of Statistics for Circle relationships

		Mean(ε =0)	Standard Dev.	Quant. at 2.5%	Quant. at 97.5%
Correlation	Pearson	-0.0019(0.000)	0.0544	-0.0919	0.0929
	Distance	0.2016(0.218)	0.0149	0.1751	0.2298
Covariance	Pearson	-0.0010	0.0277	-0.0468	0.0472
	Distance	0.1110	0.0079	0.0970	0.1263
p -value	Pearson	0.5319	0.2193	0.1506	0.9499
	Distance	0.0135	0.0073	0.0050	0.0300

Pearson correlation *r* is close to 0 which tells us that *x* and *y* are uncorrelated to each other. In contrast, distance correlation *R* tells us that x and y might be related to each other since distance correlation *R* is equal to 0.2016. Further, the mean *p* -value for the Pearson correlation test is 0.5319 giving insufficient evidence to reject the null hypothesis and we conclude that

x and y are uncorrelated. In contrast, the mean p -value for distance correlation test is 0.0135. At α level 0.1, we have sufficient evidence to reject the null hypothesis and we conclude that x and y are correlated to each other. Again, we make different conclusions when x and y are circle (nonlinearly) dependent.

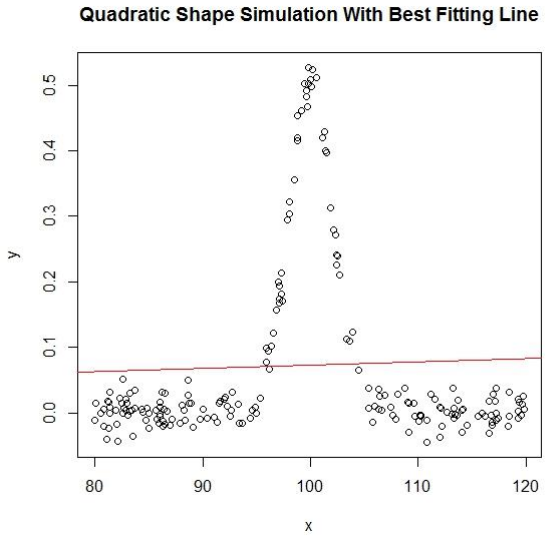
2.3.4 Peak Shape

We propose the following nonlinear model:

$$y = \frac{1}{2}\exp\left\{\frac{-(x-100)^2}{8}\right\} + \varepsilon$$

where x is sampled from uniform distribution with minimum = 80 and maximum = 120. The random error term ε is sampled from normal distribution with mean = 0 and standard deviation = 0.02 (less noise).

We randomly sample and compute the same statistics as before. Figure 2.3.4 is a sample scatter plot with best fitting line.



Below is table 2.3.4 which is the summary of statistics.

Table 2.3.4 Summary of Statistics for Peak relationships

		Mean(ε =0)	Standard Dev.	Quant. at 2.5%	Quant. at 97.5%
Correlation	Pearson	-0.0029(0.003)	0.0360	-0.0626	0.0760
	Distance	0.3575(0.365)	0.0217	0.3144	0.3976
Covariance	Pearson	-0.0047	0.0559	-0.1013	0.1098
	Distance	0.2950	0.0338	0.2356	0.3603
p -value	Pearson	0.6980	0.2027	0.2814	0.9826
	Distance	0.0050	<0.0001	0.0050	0.0050

Pearson correlation r tells us that *x* and *y* are uncorrelated to each other. In contrast, distance correlation R tells us that *x* and *y* might be related to each other since the mean distance correlation *R* is equal to 0.3575. Further, the mean *p* -value for the Pearson correlation test is 0.6980. We have insufficient evidence to reject null hypothesis and we conclude that *x* and *y* are uncorrelated. In contrast, the mean *p* -value for the distance correlation test is 0.0050. At α level 0.1, we have sufficient evidence to reject the null hypothesis and we conclude that *x* and *y* are correlated to each other. We make different conclusions when *x* and *y* are quadratic (nonlinearly) correlated.

2.3.5 Diamond Shape

We propose the following nonlinear model:

$$y_1 = \frac{2}{3}x + 2, \quad y_2 = \frac{2}{3}x - 2, \quad y_3 = -\frac{2}{3}x - 2, \quad y_4 = -\frac{2}{3}x + 2.$$

We simulate and study the data points in the area whose boundaries are the above four lines, where 'x' is sampled from uniform distribution with minimum = -3 and maximum = 3. The error term ε is sampled from normal distribution with mean = 0 and standard deviation = 0.05 (less noise). We randomly sample around 500 times to fill in the shape and calculate the same statistics as before. Figure 2.3.5 is a sample scatter plot with best fitting line.

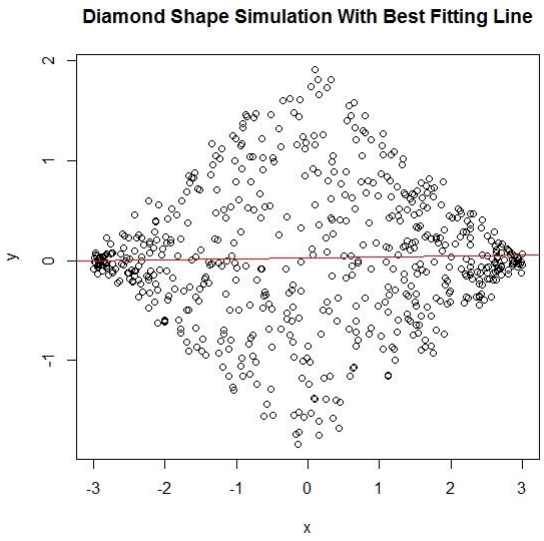


Figure 2.3.5 Diamond Shape Simulation

Below is table 2.3.5 which is the summary of statistics.

Table 2.3.5 Summary of Statistics for Diamond relationships

		Mean(ε =0)	Standard Dev.	Quant. at 2.5%	Quant. at 97.5%
Correlation	Pearson	-0.0005(0.002)	0.0362	-0.0566	0.0774
	Distance	0.2092(0.211)	0.0130	0.1850	0.2308
Covariance	Pearson	-0.0007	0.0416	-0.0662	0.0880
	Distance	0.1491	0.0109	0.1298	0.1685
p -value	Pearson	0.6366	0.2294	0.1878	0.9776
	Distance	0.0054	0.0015	0.0050	0.0100

Pearson correlation r tells us that x and y are uncorrelated to each other. In contrast, the distance correlation R tells us that x and y might be related to each other since the mean distance correlation R is equal to 0.2092. Further, the mean p -value for Pearson correlation test is 0.6366. We have insufficient evidence to reject null hypothesis and we conclude that x and y are independent. In contrast, the mean p -value for distance correlation test is 0.0054. At α level 0.1, we have sufficient evidence to reject the null hypothesis and we conclude that x and y are related to each other. We make different conclusions when x and y are diamond (nonlinearly) correlated.

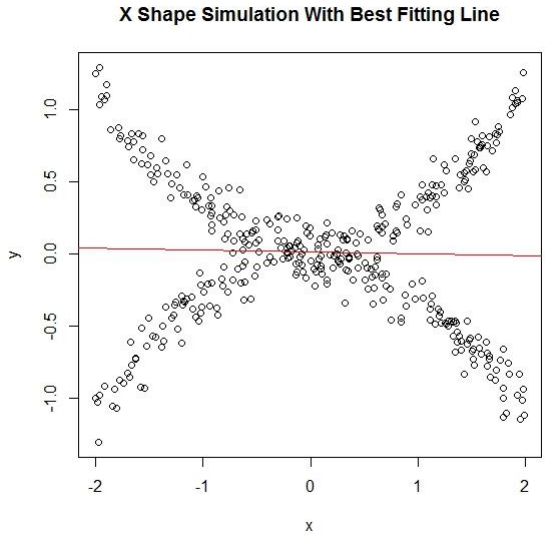
2.3.6 X Shape

We propose the following nonlinear model:

$$y = \pm(\frac{1}{4}x^2 + 0.08) + \varepsilon$$

where x is sampled from uniform distribution with minimum = -2 and maximum = 2. The error term ε is sampled from normal distribution with mean = 0 and standard deviation = 0.1.

We randomly sample around 500 times and calculate the same statistics as before. Figure 2.3.6 is a sample scatter plot with best fitting line.



Below is table 2.3.6 which is the summary of statistics.

Table 2.3.6 Summary of Statistics for X-shape relationships

		Mean(ε =0)	Standard Dev.	Quant. at 2.5%	Quant. at 97.5%
Correlation	Pearson	-0.0041(-0.001)	0.0692	-0.1356	0.1252
	Distance	0.2775(0.294)	0.0139	0.2560	0.3091
Covariance	Pearson	-0.0025	0.0414	-0.0793	0.0678
	Distance	0.1486	0.0068	0.1379	0.1616
p -value	Pearson	0.4104	0.3298	0.0058	0.9691
	Distance	0.0050	<0.0001	0.0050	0.0050

Pearson correlation r tells us that x and y are uncorrelated to each other. In contrast, the distance correlation R tells us that x and y might be related to each other since distance correlation R is equal to 0.2775. Further, the mean p -value for Pearson correlation test is 04104. We have insufficient evidence to reject null hypothesis and we conclude that x and y are independent. In contrast, the mean p -value for distance correlation test is 0.0050. At α level 0.1, we have sufficient evidence to reject the null hypothesis and we conclude that x and y are related to each other. We make different conclusions when x and y are X shape (nonlinearly) correlated.

2.3.7 Cluster Shape

We propose the following nonlinear model:

$$(y_1-2)^2+(x_1-2)^2=1,\ (y_2+2)^2+(x_2-2)^2=1,\ (y_3+2)^2+(x_3+2)^2=1,\ (y_4-2)^2+(x_4+2)^2=1.$$

We simulate and study the data points inside the four circles. 'x' is sampled from two uniform distribution with minimum = -3 and maximum = -1, minimum = 1 and maximum = 3 respectively. The error term ε is sampled from normal distribution with mean = 0 and standard deviation = 0.02 (less noise).

We randomly sample 1000 times to fill in the shape and calculate the same statistics as before. Figure 2.3.7 is a sample scatter plot with best fitting line.

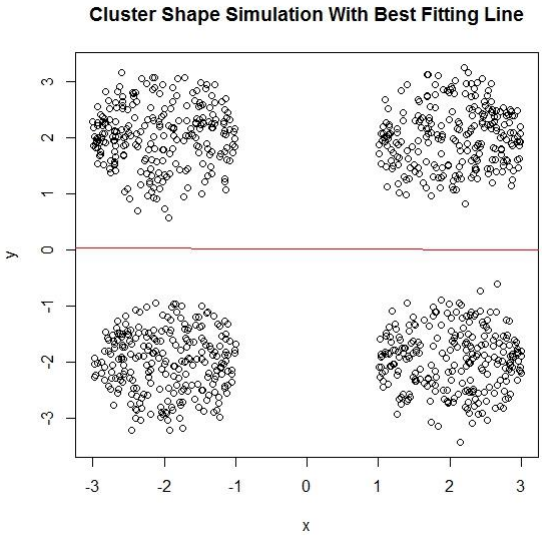


Figure 2.3.7 Cluster Shape Simulation

Below is table 2.3.7 which is the summary of statistics.

Table 2.3.7 Summary of Statistics for cluster relationships

		Mean	Standard Dev.	Quant. at 2.5%	Quant. at 97.5%
Correlation	Pearson	0.0018	0.0093	-0.0142	0.0218
	Distance	0.0283	0.0044	0.0208	0.0399
Covariance	Pearson	0.0080	0.0402	-0.0617	0.0955
	Distance	0.0496	0.0078	0.0366	0.0702
p -value	Pearson	0.8100	0.1279	0.4919	0.9884
	Distance	0.7699	0.1608	0.3548	0.9926

In contrast, both Pearson correlation and distance correlation make the same conclusion on the dependence between cluster shape simulated data x and y. The mean of Pearson correlation and the mean of distance correlation are close to 0 and both of them are smaller than 0.05 which indicate that x and y are uncorrelated to each other. Further, mean p-values for both Pearson correlation test and distance correlation test are very large. In this case, we have insufficient evidence to reject the null hypothesis and conclude that x and y are independent to each other.

2.4 CONCLUSION

If two vectors are linearly correlated or independent, we may make similar conclusion from either Pearson correlation test or distance correlation test. If two vectors are nonlinearly correlated to each other, distance correlation outperforms Pearson correlation and distance correlation could help us to detect the true relationship. Distance correlation has properties of a true dependence measure. However, distance correlation can only tell us how strongly two vectors are correlated to each other while Pearson correlation could tell us whether two vectors are positively or negatively linearly related. Therefore, in this thesis, we use both distance correlation and Pearson correlation to help us select significant variables as well as full models and do the analysis (model selection).

CHAPTER 3

EXPLORATORY DATA ANALYSIS

3.1 INTRODUCTION

Canadian crop yield threatened by climate change (especially temperature and precipitation change) is one of the most important challenges in global food security. It is quite necessary to understand how climate change impacts Canadian crop yield in order to guarantee sufficient food supply for the continuous increasing demand while sustaining the already stressed environment. In this thesis, we are going to associate crop yield with key climate factors in the last 20 to 30 years for Canadian Pairie cropping systems, in order to identify if and how climate change (especially temperature and precipitation change) is impacting crop yield. Equipped with this information, we are able to estimate the future viability of food crops in Canada. Also, we could deal with our daily consuming better and do our best to prevent food shortage in the future.

The most complete data addressing the relationship between crop yield and climate on record has been collected by Dr. Rosalind Bueckert of the Department of Plant Sciences, The University of Saskatchewan. The crop yield dataset includes yield information on 15 food crops observed in different regions within the provinces of Alberta, Saskatchewan and Manitoba, accounting for 84% of the total crop land in Canada. In this thesis, we investigate 6 different types of crops (barley, canola, flax, oats, peas and spring wheat) in Regina of Saskatchewan. Different crop yield records span over different years which are between 20 and 30 years. These

yields were recorded in the unit of bushel/acre. Also, the density of seeds planted is assumed to be constant over each year.

The climate variables were recorded daily and include maximum temperature, minimum temperature, mean temperature, rainfall, snowfall and total precipitation (in which snowfall is converted to a rain equivalent). The daily climate record in Regina spans over 100 years, which is from 1891 to 2008. We made two important modifications in the climate data of Regina. First, we create a new variable named difference temperature which is defined by the difference between maximum temperature and minimum temperature. Second, we collapse the daily data into monthly data by computing the mean in each month for each variable. For example, for minimum temperature in January 1891, we add the 31 values (31 days) up and calculate the mean minimum temperature. This mean minimum temperature is one data point of monthly data which is for January in 1891.

The growing and harvest season in Canada is from April to September while the fallow is from October to March of the next year. Despite the complexity of other factors, we show that how temperature and precipitation affect crop yield in growing season from April to September. We assume that minimum temperature and maximum temperature are linearly independent variables, so that difference temperature and mean temperature are linearly independent as well. The variables such as minimum temperature and difference temperature are not linearly independent. We cannot include two dependent variables into one model. Otherwise, the least square estimates will not be unique. Therefore, we could only include either minimum and maximum temperature or difference and mean temperature in our full model. Here, we choose difference temperature and mean temperature as well as precipitation.

3.2 BARLEY

The yield data for barley is from 1976 to 2006 with no missing data. We investigate the individual dependency of yield data and difference temperature monthly data, mean temperature monthly data and precipitation monthly data during the growing season. The relationships are plotted as below.

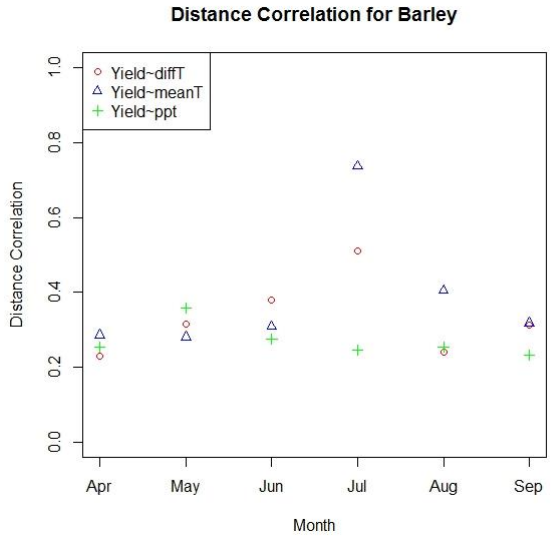


Figure 3.2.1 Distance Correlation for Barley

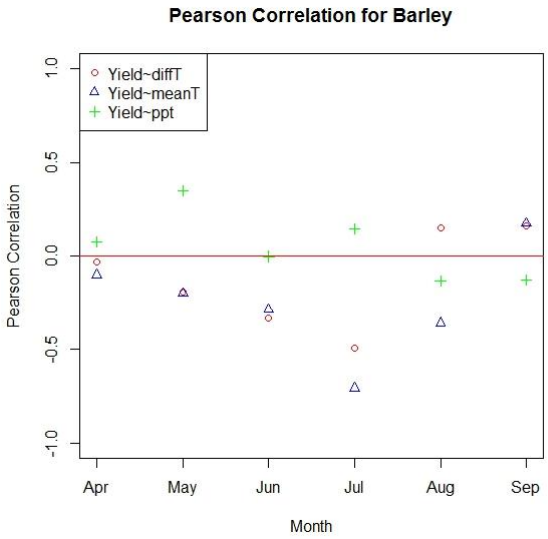


Figure 3.2.2 Pearson Correlation for Barley

Based on Figure 3.2.1, we find that diffT in July, meanT in July and ppt in May have the relatively biggest distance correlation R. This indicates that they may significantly affect barley yield. In Figure 3.2.2, we find Pearson correlation r for diffT in July and meanT in July are negative while r is positive for ppt in May. This indicates that barley yield is negatively correlated to diffT and meanT in July and positively correlated to ppt in May. We may observe these relationships from the scatter plots with best fitting line shown as below.

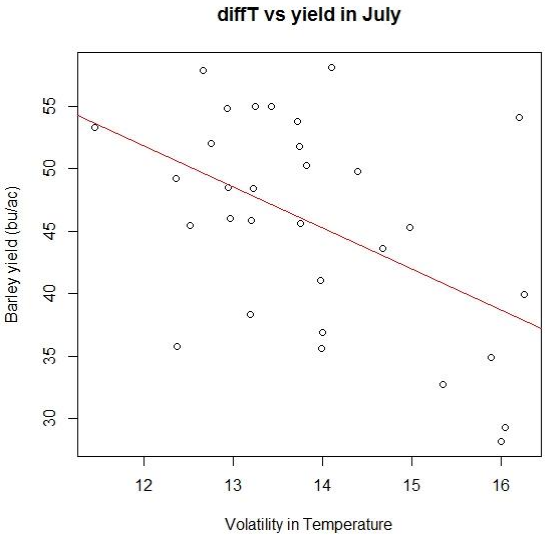


Figure 3.2.3 Barley yield vs diffT in July

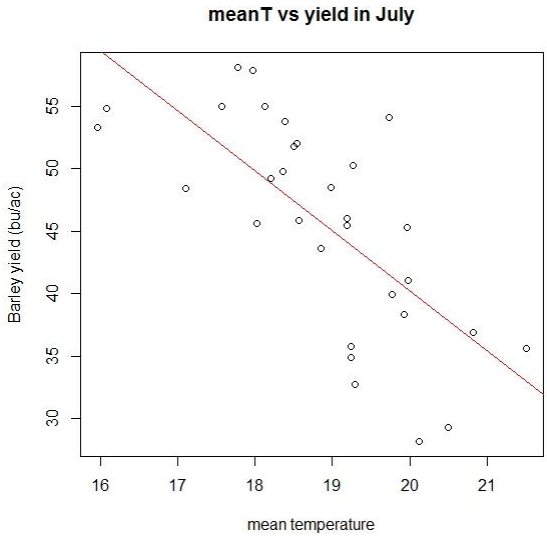


Figure 3.2.4 Barley yield vs meanT in July

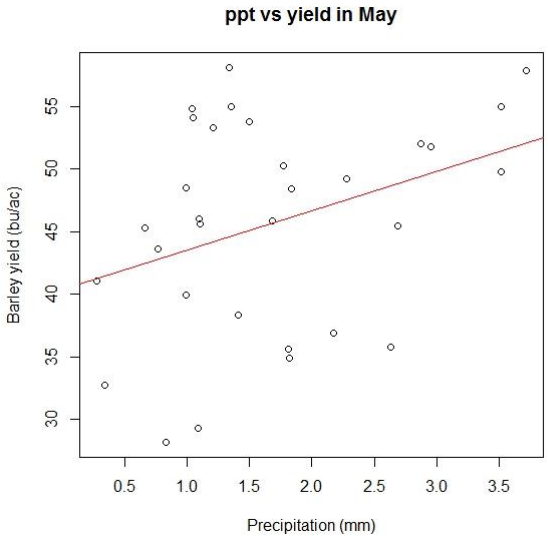


Figure 3.2.5 Barley yield vs ppt in May

Further, we calculate both distance correlation R and Pearson correlation r for all three variables in different months. We perform both distance correlation test and Pearson correlation test and obtain p-values for all of them. These statistics are presented in Table 3.2.1.

Table 3.2.1 Summary of Statistics for Barley Yield

		Barley Yield	
		Distance Cor. (p -value)	Pearson Cor. (p -value)
Difference Temperature (diffT)	April	0.2302 (0.8950)	-0.0345 (0.8537)
	May	0.3139 (0.3500)	-0.1951 (0.2929)
	June	0.3806 (0.0750)	-0.3333 (0.0669)
	July	0.5107 (0.0150)	-0.4925 (0.0049)
	August	0.2412 (0.8200)	0.1482 (0.4264)
	September	0.3120 (0.3100)	0.1582 (0.3953)
Mean Temperature (meanT)	April	0.2856 (0.5050)	-0.1009 (0.5893)
	May	0.2809 (0.5300)	-0.2019 (0.2760)
	June	0.3084 (0.3800)	-0.2856 (0.1192)
	July	0.7370 (0.0050)	-0.7060 (9.124e-6)
	August	0.4047 (0.0650)	-0.3585 (0.0477)
	September	0.3174 (0.2800)	0.1741 (0.3489)
Precipitation (ppt)	April	0.2540 (0.6300)	0.0727 (0.6974)
	May	0.3582 (0.1150)	0.3461 (0.0565)
	June	0.2742 (0.4900)	-0.0035 (0.9850)
	July	0.2450 (0.8100)	0.1442 (0.4389)
	August	0.2527 (0.6850)	-0.1340 (0.4723)
	September	0.2318 (0.7900)	-0.1281 (0.4921)

Based on the numerical results in the above table, both distance correlation R and Pearson correlation r of diffT in July, meanT in July and ppt in May are the biggest values relatively. This result is consistent with our observations from distance correlation and Pearson correlation scatter plots. By checking p -values in both Pearson correlation test and Distance correlation test, we find diffT in June and July, meanT in July and August and ppt in May are significant at α level 0.1. This implies that barley yield is significantly impacted by these factors and all of them should be included into our full model.

3.3 CANOLA

The yield data for canola is from 1980 to 2006 with no missing data. We investigate the individual dependency of yield data the same as barley. The relationships are plotted as below.

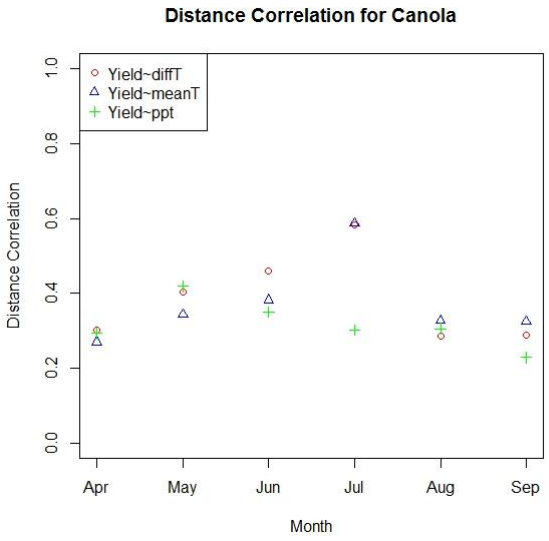


Figure 3.3.1 Distance Correlation for Canola

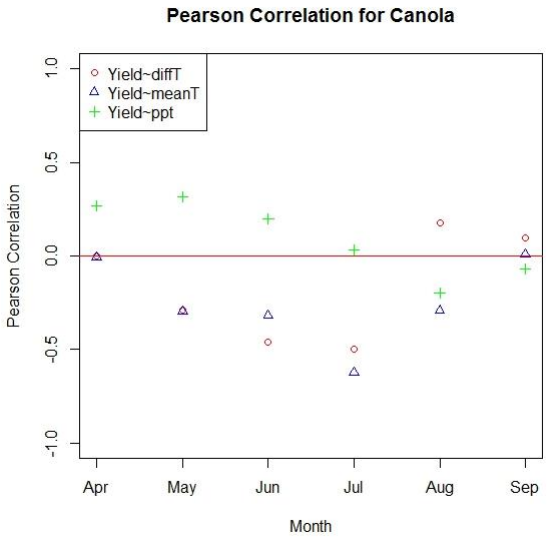


Figure 3.3.2 Pearson Correlation for Canola

Based on Figure 3.3.1, we find that diffT in July, meanT in July and ppt in May have the biggest distance correlation *R* relatively. This indicates that climate in these months may significantly affect canola yield. In Figure 3.3.2, we find Pearson correlation r for diffT in July and meanT in July are negative while Pearson correlation *r* is positive for ppt in May. This indicates that canola yield is negatively correlated to diffT and meanT in July and positively correlated to ppt in May. We may also observe these relationships from the scatter plots with best fitting line shown as below.

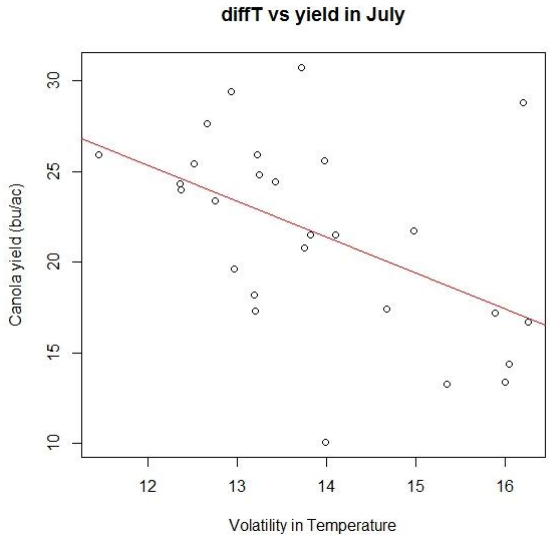


Figure 3.3.3 Canola Yield vs diffT in July

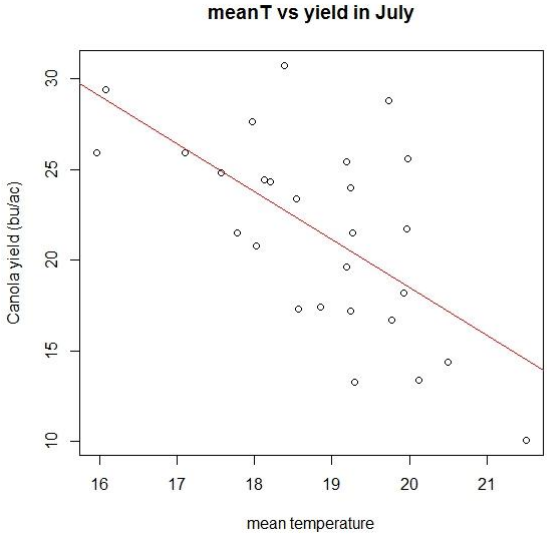


Figure 3.3.4 Canola Yield vs meanT in July

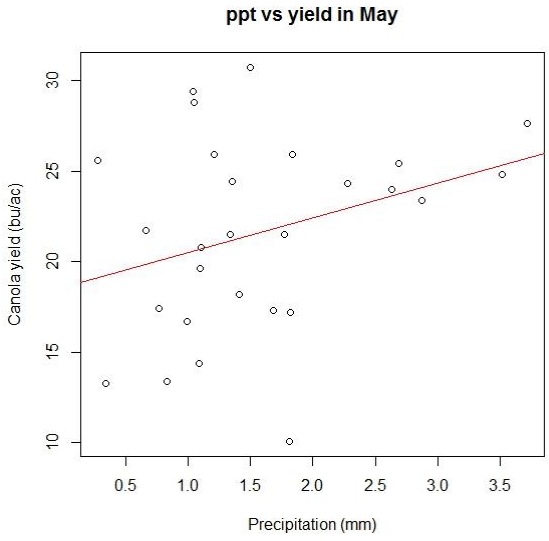


Figure 3.3.5 Canola Yield vs ppt in May

Further, we calculate both distance correlation R and Pearson correlation r for all three variables in different months. We perform both distance correlation test and Pearson correlation test and obtain p-values for all of them. These statistics are presented in Table 3.3.1.

Table 3.3.1 Summary of Statistics for Canola Yield

		Canola Yield	
		Distance Cor. (p -value)	Pearson Cor. (p -value)
Difference Temperature (diffT)	April	0.3019 (0.5650)	-0.0028 (0.9890)
	May	0.4021 (0.1550)	-0.2899 (0.1424)
	June	0.4588 (0.0550)	-0.4592 (0.0160)
	July	0.5833 (0.0050)	-0.4974 (0.0083)
	August	0.2865 (0.6200)	0.1740 (0.3853)
	September	0.2883 (0.6750)	0.0972 (0.6300)
Mean Temperature (meanT)	April	0.2702 (0.7800)	-0.0066 (0.9739)
	May	0.3440 (0.3450)	-0.2969 (0.1326)
	June	0.3818 (0.2350)	-0.3199 (0.1038)
	July	0.5874 (0.0100)	-0.6253 (0.0005)
	August	0.3271 (0.3350)	-0.2923 (0.1389)
	September	0.3238 (0.4050)	0.0078 (0.9692)
Precipitation (ppt)	April	0.2923 (0.6250)	0.2667 (0.1788)
	May	0.4182 (0.1100)	0.3150 (0.1095)
	June	0.3496 (0.2400)	0.1975 (0.3234)
	July	0.3009 (0.5750)	0.0310 (0.8782)
	August	0.3054 (0.4450)	-0.2005 (0.3160)
	September	0.2281 (0.9300)	-0.0695 (0.7305)

Based on the numerical results in the above table, both distance correlation R and Pearson correlation r of diffT in July, meanT in July and ppt in May are the biggest values relatively. This result is consistent with our observations from distance correlation and Pearson correlation scatter plots. By checking p -values in both Pearson correlation test and Distance correlation test, we find diffT in June and July, meanT in July are significant at α level 0.1. The p -value of Pearson correlation test for meanT in June is equal to 0.1038 and the p -value of Pearson correlation test for ppt in May is 0.1095. Although the two p -values are slightly bigger than 0.1, we still consider the two factors as significant factors. Therefore, canola yield is significantly impacted by these 5 factors and all of them should be included into our full model.

3.4 FLAX

The yield data for flax is from 1976 to 2006 with no missing data. We investigate the individual dependency of yield data the same as barley. The relationships are plotted as below.

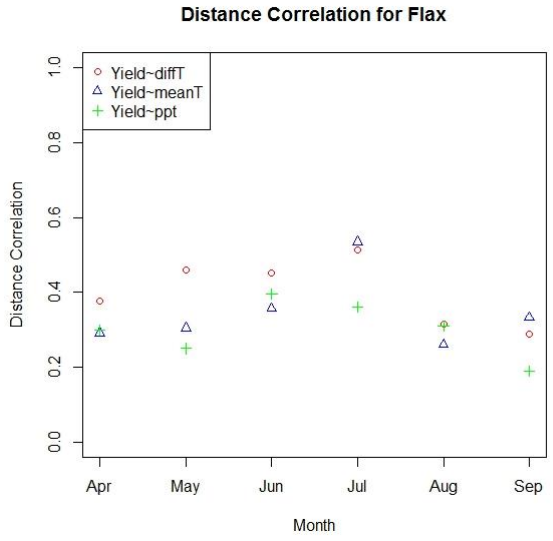


Figure 3.4.1 Distance Correlation for Flax

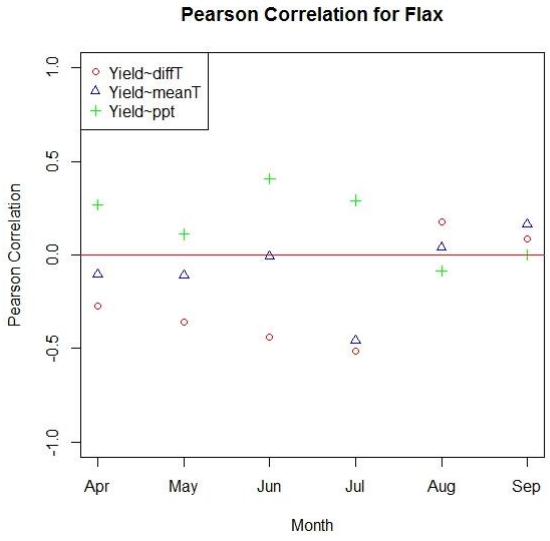


Figure 3.4.2 Pearson Correlation for Flax

Based on Figure 3.4.1, we find that diffT in July, meanT in July and ppt in June have the biggest distance correlation *R* relatively. This indicates that climate in these months may significantly affect flax yield. In Figure 3.4.2, we find Pearson correlation r for diffT in July and meanT in July are negative while Pearson correlation r is positive for ppt in June. This indicates that flax yield is negatively correlated to diffT and meanT in July and positively correlated to ppt in June. We may observe these relationships from the scatter plots with best fitting line shown as below.

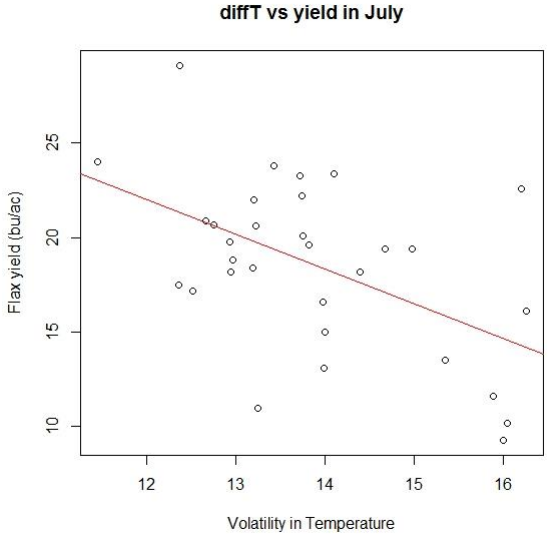


Figure 3.4.3 Flax Yield vs diffT in July

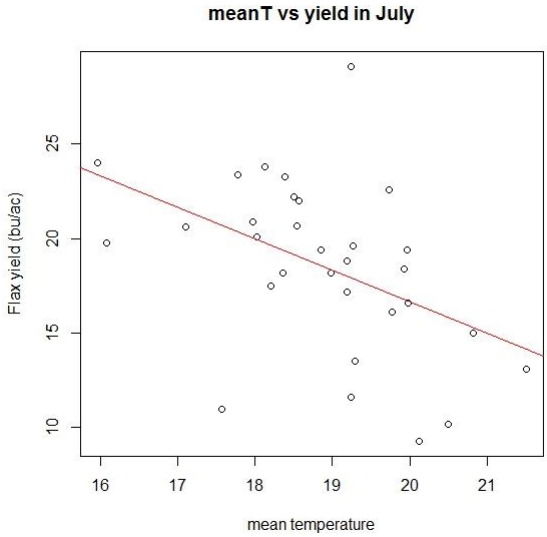


Figure 3.4.4 Flax Yield vs meanT in July

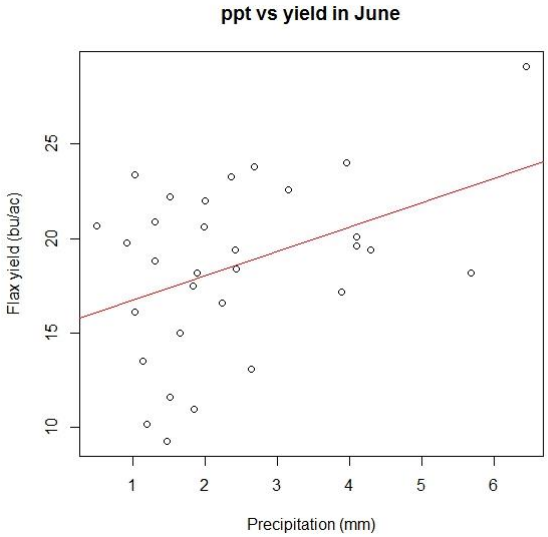


Figure 3.4.5 Flax Yield vs ppt in June

Further, we calculate both distance correlation R and Pearson correlation r for all three variables in different months. We perform both distance correlation test and Pearson correlation test and obtain p-values for all of them. These statistics are presented in table 3.4.1.

Table 3.4.1 Summary of Statistics for Flax Yield

		Flax Yield	
		Distance Cor. (p -value)	Pearson Cor. (p -value)
Difference Temperature (diffT)	April	0.3753 (0.1450)	-0.2755 (0.1335)
	May	0.4587 (0.0350)	-0.3601 (0.0466)
	June	0.4510 (0.0400)	-0.4405 (0.0131)
	July	0.5124 (0.0250)	-0.5150 (0.0030)
	August	0.3139 (0.3200)	0.1758 (0.3443)
	September	0.2895 (0.5950)	0.0874 (0.6401)
Mean Temperature (meanT)	April	0.2911 (0.4800)	-0.1059 (0.5706)
	May	0.3047 (0.3900)	-0.1105 (0.5542)
	June	0.3574 (0.2200)	-0.0087 (0.9630)
	July	0.5344 (0.0150)	-0.4578 (0.0096)
	August	0.2599 (0.7500)	0.0417 (0.8238)
	September	0.3322 (0.2850)	0.1630 (0.3809)
Precipitation (ppt)	April	0.2982 (0.4550)	0.2686 (0.1440)
	May	0.2500 (0.8200)	0.1096 (0.5571)
	June	0.3954 (0.0650)	0.4048 (0.0239)
	July	0.3602 (0.1750)	0.2912 (0.1119)
	August	0.3091 (0.2950)	-0.0866 (0.6432)
	September	0.1889 (0.9950)	0.0003 (0.9989)

Based on the numerical results in the above table, both distance correlation R and Pearson correlation r of diffT in July, meanT in July and ppt in June are the biggest values relatively. This result is consistent with our observations from distance correlation and Pearson correlation scatter plots. By checking p -values in both Pearson correlation test and Distance correlation test, we find diffT in May, June and July, meanT in July and ppt in June are significant at α level 0.1. This implies that flax yield is significantly impacted by these 5 factors and all of them should be included into our full model.

3.5 OATS

The yield data for oats is from 1976 to 2006 with no missing data. We investigate the individual dependency of yield data the same as barley. The relationships are plotted as below.

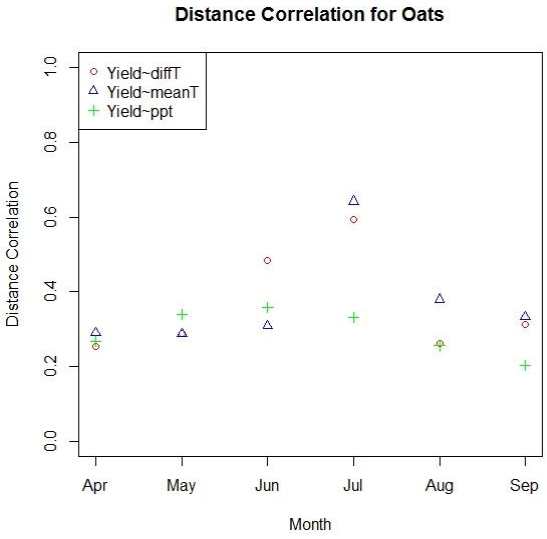


Figure 3.5.1 Distance Correlation for Oats

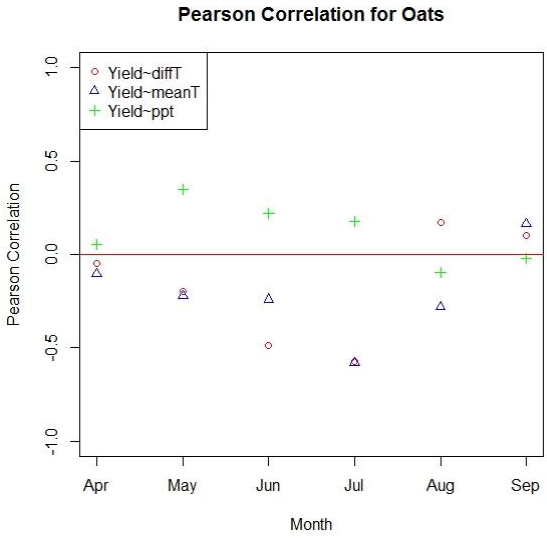
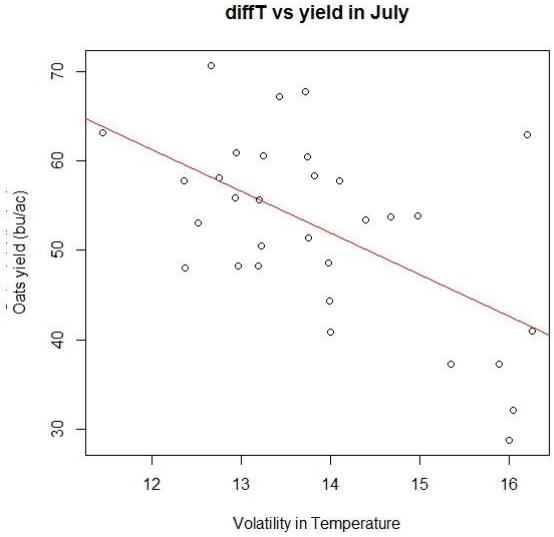
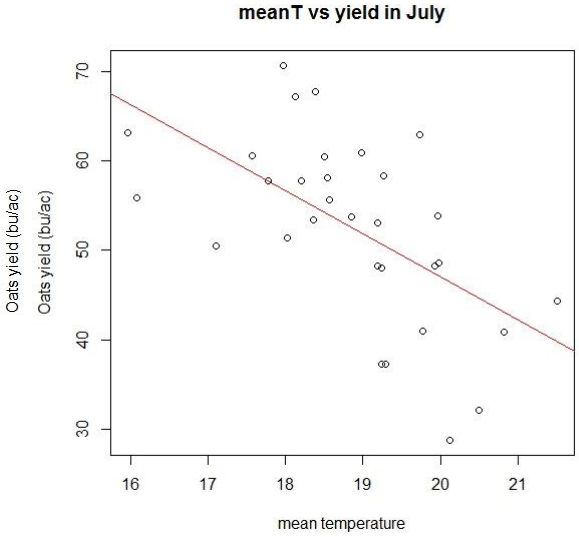
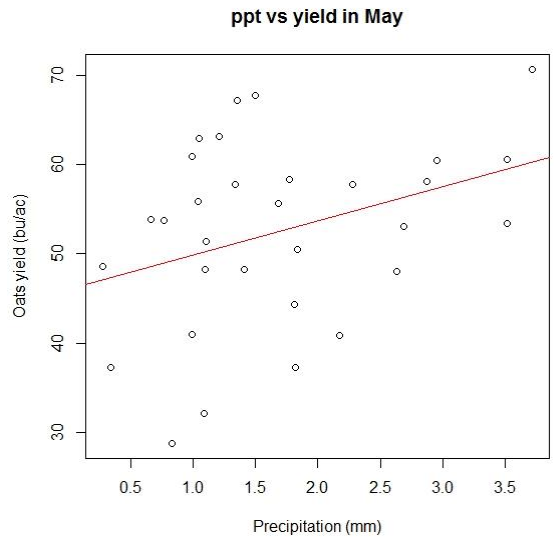


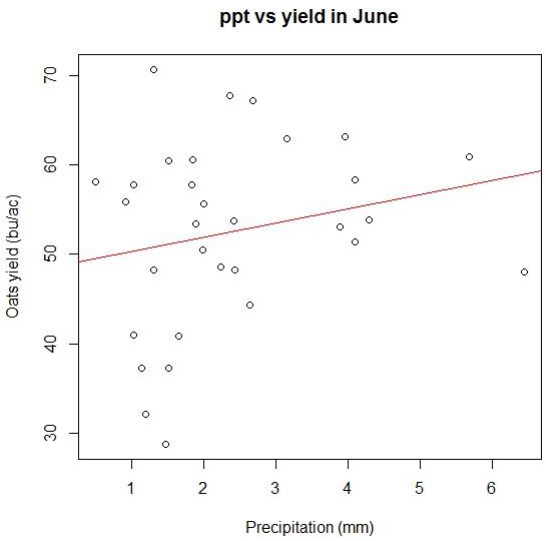
Figure 3.5.2 Pearson Correlation for Oats

Based on Figure 3.5.1, we find that diffT in July, meanT in July and ppt in June have the biggest distance correlation *R* relatively. This indicates that climate in these months may significantly affect oats yield. In Figure 3.5.2, we have the same conclusion as distance correlation *R* except Pearson correlation *r* of ppt in May which is bigger than *r* of ppt in June. The Pearson correlation r for diffT in July and meanT in July are negative while Pearson correlation r is positive for ppt in both May and June. This indicates that oats yield is negatively correlated to diffT and meanT in July and positively correlated to ppt in May and June. We may observe these relationships from the scatter plots with best fitting line shown as below.









Further, we calculate both distance correlation R and Pearson correlation r for all three variables in different months. We perform both distance correlation test and Pearson correlation test and obtain p-values for all of them. These statistics are presented in Table 3.5.1.

Table 3.5.1 Summary of Statistics for Oats Yield

		Oats Yield	
		Distance Cor. (p -value)	Pearson Cor. (p -value)
Difference Temperature (diffT)	April	0.2538 (0.8800)	-0.0503 (0.7882)
	May	0.2892 (0.5950)	-0.2009 (0.2785)
	June	0.4832 (0.0300)	-0.4879 (0.0054)
	July	0.5935 (0.0050)	-0.5743 (0.0007)
	August	0.2615 (0.6300)	0.1713 (0.3569)
	September	0.3131 (0.4200)	0.1012 (0.5855)
Mean Temperature (meanT)	April	0.2905 (0.5600)	-0.1030 (0.5815)
	May	0.2869 (0.5550)	-0.2215 (0.2310)
	June	0.3080 (0.6100)	-0.2408 (0.1919)
	July	0.6415 (0.0050)	-0.5814 (0.0006)
	August	0.3794 (0.1100)	-0.2796 (0.1277)
	September	0.3325 (0.2250)	0.1630 (0.3810)
Precipitation (ppt)	April	0.2662 (0.7150)	0.0545 (0.7711)
	May	0.3387 (0.2300)	0.3454 (0.0570)
	June	0.3581 (0.1700)	0.2206 (0.2330)
	July	0.3312 (0.2850)	0.1775 (0.3395)
	August	0.2549 (0.7250)	-0.0951 (0.6107)
	September	0.2035 (0.9650)	-0.0236 (0.8996)

Based on the numerical results in the above table, both distance correlation R and Pearson correlation r of diffT and meanT in July are the biggest values relatively. However, distance correlation R tells ppt in June is the biggest while Pearson correlation r indicates ppt in May is the biggest. This result is consistent with our observations from distance correlation and Pearson correlation scatter plots. By checking p-values in both Pearson correlation test and Distance correlation test, we find diffT in June and July, meanT in July and ppt in May are significant at α level 0.1. We don't think ppt in June is significant because p-value of neither distance correlation test nor Pearson Correlation test is smaller than 0.1. Therefore, oats yield is significantly affected by 4 factors except ppt in June and all of 4 factors should be included into our full model.

3.6 PEA

The yield data for pea is from 1987 to 2006 with no missing data. We investigate the individual dependency of yield data the same as barley. The relationships are plotted as below.

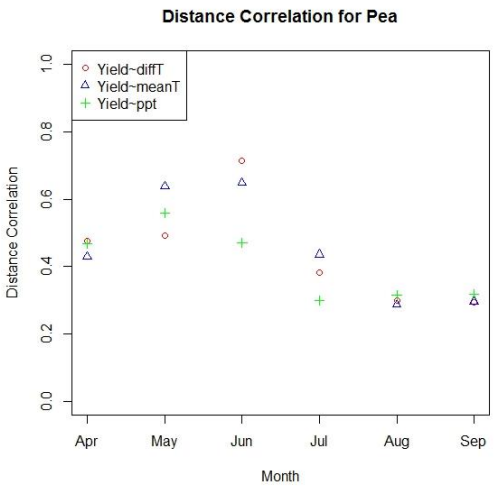


Figure 3.6.1 Distance Correlation for Pea

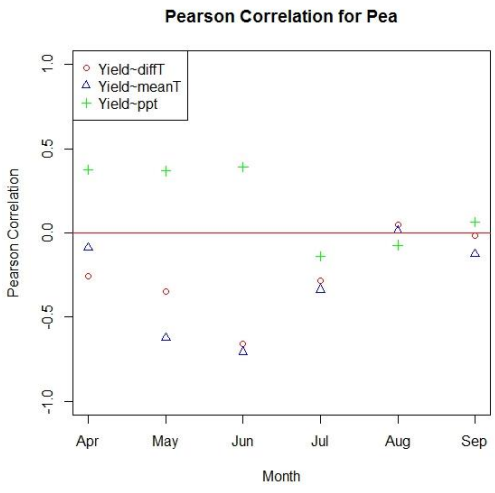
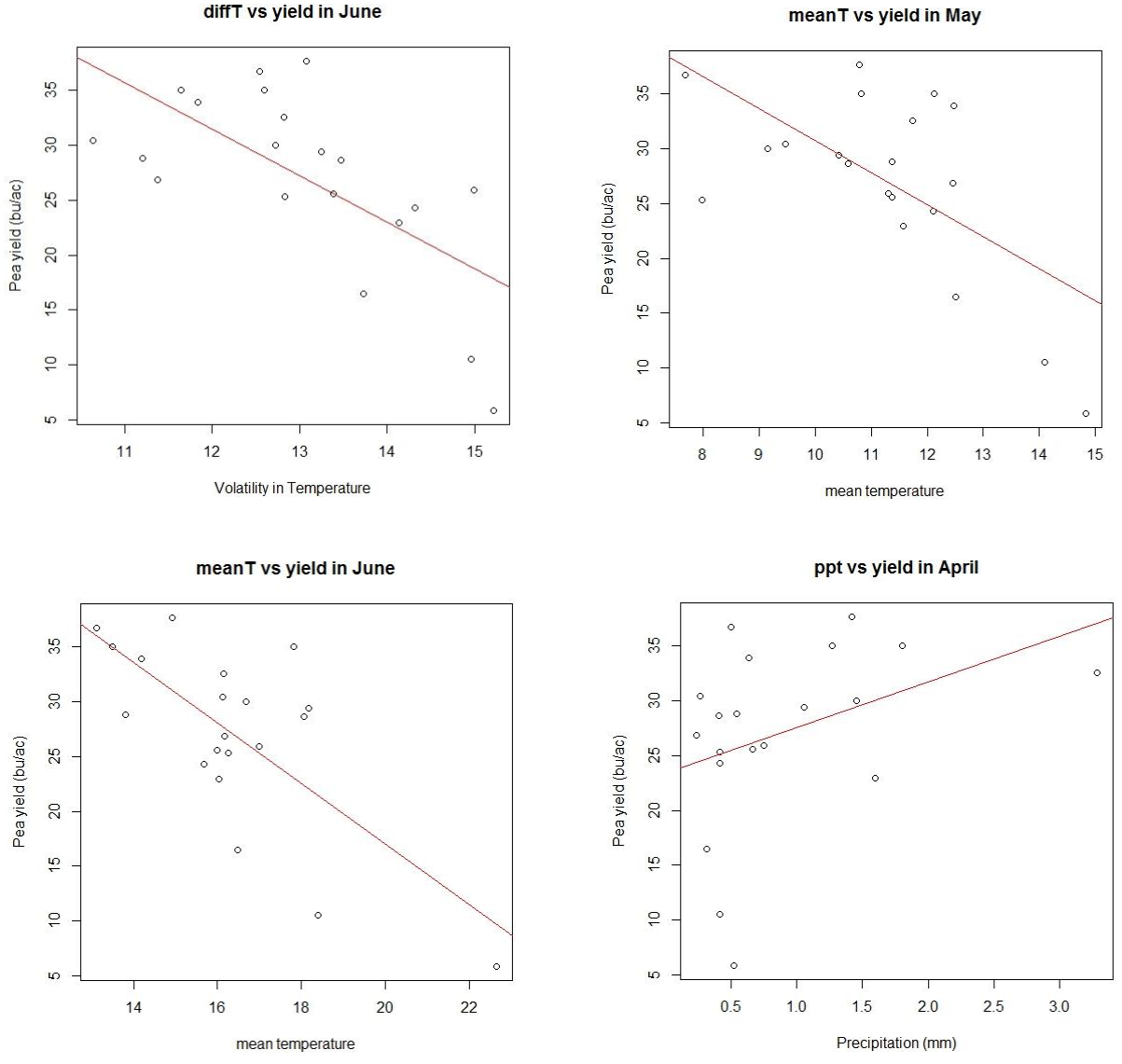
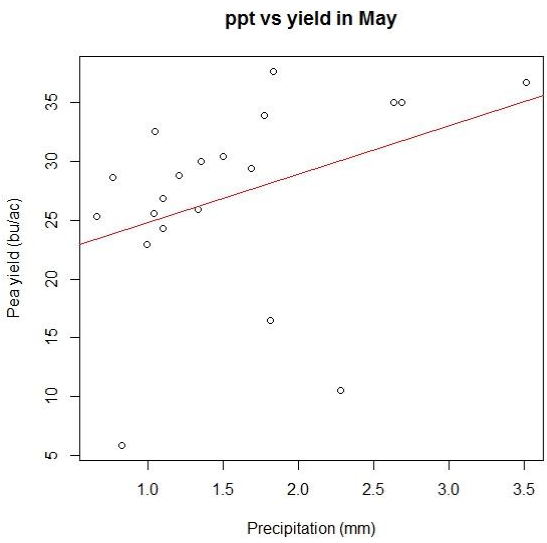
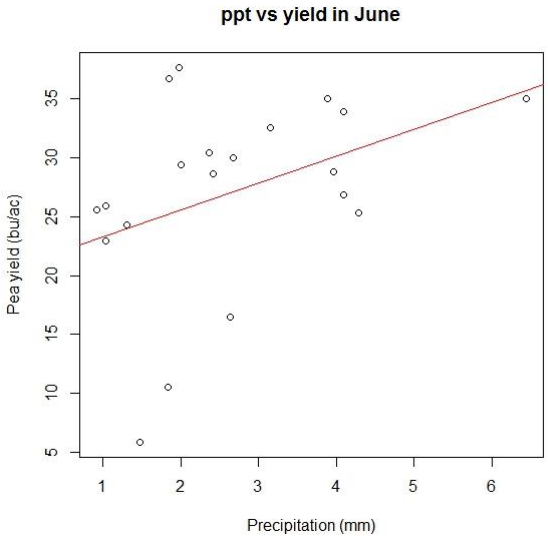


Figure 3.6.2 Pearson Correlation for Pea

Based on Figure 3.6.1, we find that diffT in June and ppt in May have the biggest distance correlation R relatively while meanT in May and June have almost the same biggest value R. This indicates that climate in these months may significantly affect pea yield. In Figure 3.6.2, we find Pearson correlation r for ppt in April, May and June are almost the same and positive. Pearson correlation r for diffT in June and meanT in May and June are negative. These imply how pea yield is associated with climate in these months. We could observe these relationships from the scatter plots with best fitting line shown as below.







Further, we calculate both distance correlation R and Pearson correlation r for all three variables in different months. We perform both distance correlation test and Pearson correlation test and obtain p -values for all of them. These statistics are presented in Table 3.6.1.

Based on the numerical results in the above table, we find that diffT in June, meanT in May and June and ppt in May have the relatively biggest distance correlation R. If we check their Pearson correlation r, we could make a similar conclusion. The only difference is that Pearson correlation r also suggests r of ppt in April and June is almost the same as ppt in May. Further, by checking p -values in both Pearson correlation test and Distance correlation test, we find diffT in June, meanT in May and June and ppt in April, May and June are significant at α level 0.1. This implies that pea yield is significantly impacted by these 6 factors and all of them should be included into our full model.

Table 3.6.1 Summary of Statistics for Pea Yield

		Pea Yield	
		Distance Cor. (p -value)	Pearson Cor. (p -value)
Difference Temperature (diffT)	April	0.4769 (0.1200)	-0.2565 (0.2750)
	May	0.4927 (0.1200)	-0.3486 (0.1320)
	June	0.7140 (0.0050)	-0.6613 (0.0015)
	July	0.3827 (0.4000)	-0.2857 (0.2221)
	August	0.2995 (0.7900)	0.0489 (0.8377)
	September	0.2938 (0.9400)	-0.0158 (0.9473)
Mean Temperature (meanT)	April	0.4296 (0.2100)	-0.0871 (0.7150)
	May	0.6381 (0.0250)	-0.6238 (0.0033)
	June	0.6494 (0.0150)	-0.7063 (0.0005)
	July	0.4362 (0.2150)	-0.3379 (0.1451)
	August	0.2872 (0.9300)	0.0125 (0.9584)
	September	0.2963 (0.8650)	-0.1244 (0.6013)
Precipitation (ppt)	April	0.4677 (0.0900)	0.3734 (0.1048)
	May	0.5586 (0.0250)	0.3666 (0.1118)
	June	0.4706 (0.1100)	0.3903 (0.0889)
	July	0.2991 (0.8850)	-0.1415 (0.5517)
	August	0.3146 (0.8150)	-0.0731 (0.7595)
	September	0.3164 (0.8000)	0.0628 (0.7525)

2.7 SPRING WHEAT

The yield data for spring wheat is from 1976 to 2006 with no missing data. We investigate the individual dependency of yield data the same as barley. The relationships are plotted as below.

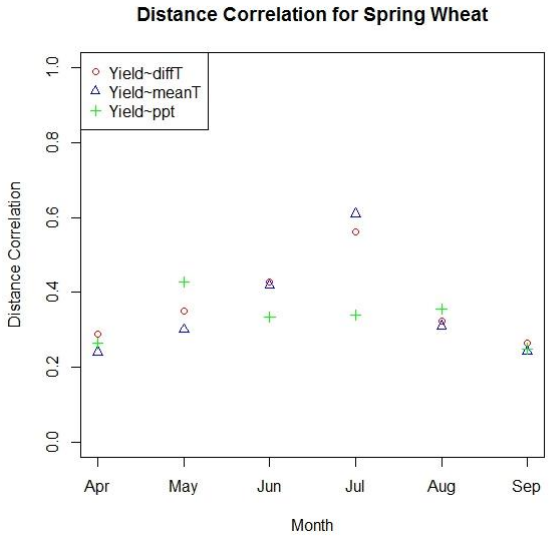


Figure 3.7.1 Distance Correlation for SWheat

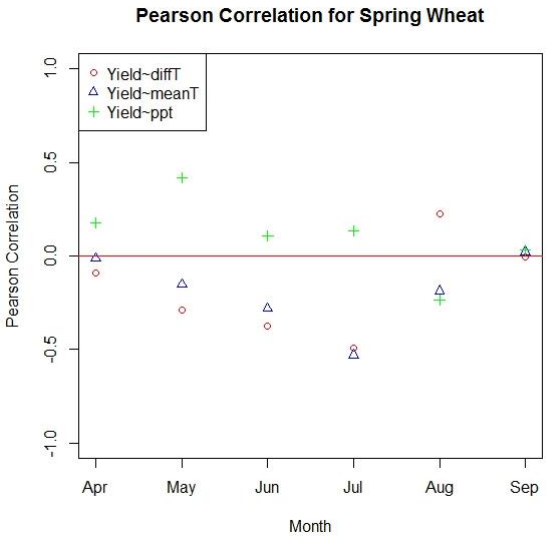


Figure 3.7.2 Pearson Correlation for SWheat

Based on Figure 3.7.1, we find that diffT in July, meanT in July and ppt in May have the biggest distance correlation *R* relatively. This indicates that climate in these months may significantly affect spring wheat yield. In Figure 3.7.2, we find Pearson correlation r for diffT in July and meanT in July are negative while Pearson correlation *r* is positive for ppt in May. This indicates that spring wheat yield is negatively correlated to diffT and meanT in July and positively correlated to ppt in May. We may observe these relationships from the scatter plots with best fitting line shown as below.

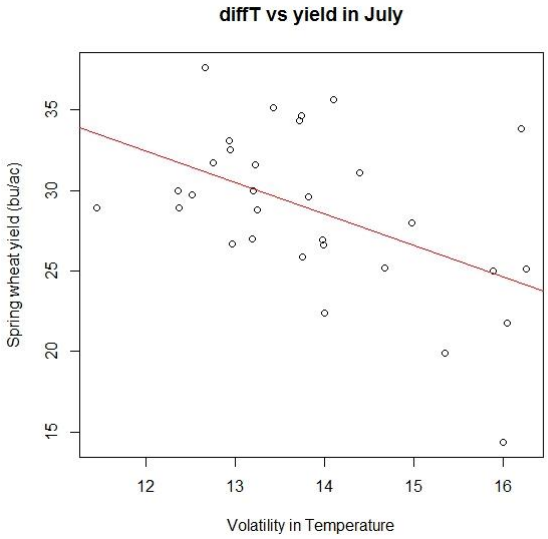


Figure 3.7.3 SWheat Yield vs diffT in July

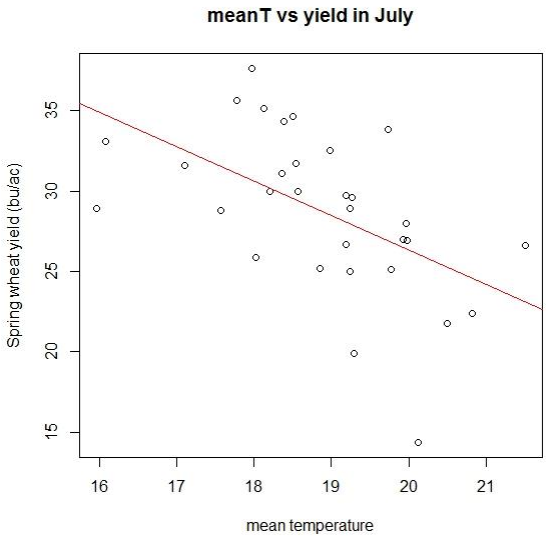


Figure 3.7.4 SWheat Yield vs meanT in July

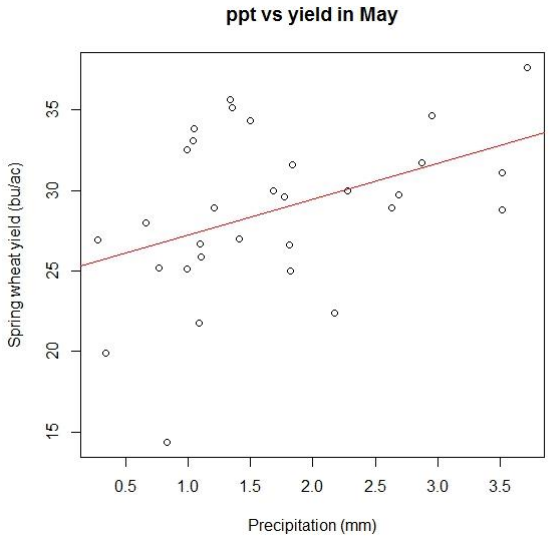


Figure 3.7.5 SWheat Yield vs ppt in May

Further, we calculate both distance correlation R and Pearson correlation r for all three variables in different months. We perform both distance correlation test and Pearson correlation test and obtain p-values for all of them. These statistics are presented in Table 3.7.1.

Based on the numerical results in the above table, both distance correlation R and Pearson correlation r of diffT in July, meanT in July and ppt in May are the biggest values relatively.

This result is consistent with our observations from distance correlation and Pearson correlation scatter plots. By checking p-values in both Pearson correlation test and distance correlation test, we find diffT in June and July, meanT in June and July and ppt in May are significant at α level 0.1. This implies that spring wheat yield is significantly impacted by these 5 factors and all of them should be included into our full model.

Table 3.7.1 Summary of Statistics for Spring Wheat Yield

		Spring Wheat Yield	
		Distance Cor. (p-value)	Pearson Cor. (p-value)
Difference Temperature (diffT)	April	0.2885 (0.6300)	-0.0894 (0.6324)
	May	0.3509 (0.2950)	-0.2873 (0.1170)
	June	0.4284 (0.0650)	-0.3751 (0.0376)
	July	0.5609 (0.0050)	-0.4950 (0.0046)
	August	0.3228 (0.3400)	0.2252 (0.2233)
	September	0.2651 (0.8400)	-0.0079 (0.9662)
Mean Temperature (meanT)	April	0.2400 (0.9250)	-0.0145 (0.9385)
	May	0.3003 (0.4900)	-0.1528 (0.4120)
	June	0.4187 (0.1000)	-0.2816 (0.1249)
	July	0.6095 (0.0050)	-0.5299 (0.0022)
	August	0.3096 (0.3950)	-0.1871 (0.3134)
	September	0.2418 (0.8550)	0.0219 (0.9071)
Precipitation (ppt)	April	0.2629 (0.7100)	0.2629 (0.3392)
	May	0.4273 (0.0400)	0.4153 (0.0202)
	June	0.3340 (0.2750)	0.1071 (0.5663)
	July	0.3383 (0.3000)	0.1328 (0.4762)
	August	0.3543 (0.2450)	-0.2345 (0.2042)
	September	0.2493 (0.8050)	0.0319 (0.8647)

CHAPTER 4

ANALYSIS AND DISCUSSION

4.1 INTRODUCTION

In this chapter, we use both distance correlation and Pearson correlation to help us select significant variables. We analyze 6 different crops which are barley, canola, flax, pea, oats and spring wheat. For each crop, we start with two full models. The first full model (Full 1) includes significant variables only. The second full model (Full 2) includes not only significant variables but also their corresponding interaction terms. Based on Effect Heredity Principle, if both main effects are not significant, their interaction should not be significant, so we will not consider interaction terms if both main effects are not significant. Here, we take barley as an example for Full 2. We include ppt in May as a significant variable, we also include its corresponding interaction terms i.e. ppt in May*diffT in May and ppt in May*meanT in May. Since we introduce two new variables diffT in May and meanT in May in the interaction terms, we have to include diffT in May and meanT in May in our full 2 as well, analogous to other significant variables. Further, we also consider cumulative precipitation in February and March (fallow months) and include this term in both Full 1 and Full 2.

Once the two full models are determined, we perform backward selection, forward selection, stepwise selection from null, and stepwise selection from full on both of them. We take consideration of various factors such as RMSE, RSS, AIC, BIC, R square, number of parameters and so on and select the best model from each starting model. Then we compare the two best

models to get the final best model for each crop. Further, we use several statistical methods to check the goodness of fit of our best model (Q-Q plot, residual plots etc). Finally, we have 6 best models for 6 different crops and we will interpret them in details as well. Our cutoff line for p-value is 0.1which is α level.

4.2 BARLEY

From Table 3.2.1, we find that p-values of either distance correlation test or Pearson correlation test for difference temperature in June and July, mean temperature in July and August and precipitation in May are smaller than 0.1. This indicates that all these five variables diffT in June, diffT in July, meanT in July, meanT in August and ppt in May are significant variables. Based on the definition for Full 1, we include these significant variables as well as the cumulative precipitation in February and March. The **Full 1** is

$$yield = \widehat{\beta_0} + \widehat{\beta_1} * diffT\ in\ Jun + \widehat{\beta_2} * diffT\ in\ Jul + \widehat{\beta_3} * meanT\ in\ Jul$$

$$+\widehat{\beta_4} * meanT\ in\ Aug + \widehat{\beta_5} * ppt\ in\ May + \widehat{\beta_6} * ppt\ in\ Feb\ and\ Mar$$

where 'yield' is barley yield.

Also, based on the definition for full model 2 in the introduction, we include significant variables and interaction terms as well as the cumulative precipitation in February and March. The **Full 2** is

$$yield = \widehat{\beta_0} + \widehat{\beta_1} * diffT\ in\ May + \widehat{\beta_2} * diffT\ in\ Jun + \widehat{\beta_3} * diffT\ in\ Jul$$

$$+\widehat{\beta_4} * meanT\ in\ May + \widehat{\beta_5} * meanT\ in\ Jul + \widehat{\beta_6} * meanT\ in\ Aug + \widehat{\beta_7} * ppt\ in\ May$$

$$+\widehat{\beta_8} * ppt\ in\ Jun + \widehat{\beta_9} * ppt\ in\ Jul + \widehat{\beta_{10}} * ppt\ in\ Aug + \widehat{\beta_{11}} * ppt\ in\ Feb\ and\ Mar$$

$$+\widehat{\beta_{12}} * diffT\ in\ May * ppt\ in\ May + \widehat{\beta_{13}} * diffT\ in\ Jun * ppt\ in\ Jun$$

$$+\widehat{\beta_{14}} * diffT\ in\ Jul * ppt\ in\ July + \widehat{\beta_{15}} * meanT\ in\ May * ppt\ in\ May$$

$$+\widehat{\beta_{16}} * meanT\ in\ Jul * ppt\ in\ Jul + \widehat{\beta_{17}} * meanT\ in\ Aug * ppt\ in\ Aug$$

For Full 1, we use **backward selection** method first, which returns

$$yield = 127.557 - 4.510 * \text{meanT in July} + 1.899 * \text{ppt in May} \qquad (4.2.1)$$

Then, we use **forward selection**, **stepwise selection** starts from **full** and **stepwise selection** starts from **null**. All the 4 methods return us the same best model for Full 1. Table 4.2.1 is the summary of this model (Parm #: Number of parameters; df: Degree freedom; RMSE: Root of mean square error; R-Sq: R Squared).

Table 4.2.1 Summary of Model Selection of Full 1 for Barley

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.2.1)	3	28	5.935	986.36	113.26	117.56	0.5406

For Full 2, we use **backward selection** method first. This returns us the following model.

$$yield = 164.1966 + 5.3544 * \text{ppt in Feb and Mar} - 2.278 * \text{diffT in Jul}$$

$$-10.3178 * \text{diffT in Jun} + 3.3696 * \text{diffT in May} - 0.6491 * \text{meanT in Jul}$$

$$-5.0044 * \text{ppt in Aug} + 12.0089 * \text{ppt in Jul} - 47.5186 * \text{ppt in Jun}$$

$$+24.0436 * \text{ppt in May} + 2.7318 * \text{diffT in Jul} * \text{ppt in Jul}$$

$$+3.8819 * \text{diffT in June} * \text{ppt in Jun} - 2.4860 * \text{meanT in Jul} * \text{ppt in Jul}$$

$$-1.4537 * \text{diffT in May} * \text{ppt in May} \qquad (4.2.2)$$

which is the same model we get using stepwise selection starting from the full model.

Then, we use **forward selection** method. This returns us the following model.

$$yield = 149.9258 - 4.928 * \text{meanT in Jul} - 5.9741 * \text{ppt in May}$$

$$-2.3793 * \text{meanT in May} \; - 2.0979 * \text{ppt in Aug} + 1.0559 * \text{diffT in May}$$

$$+0.7099 * \text{ppt in May} * \text{meanT in May} \qquad (4.2.3)$$

which is the same model we get using stepwise selection starting from the null model.

We have two options from Full 2 and we need to determine which one is better. Table 4.2.2 is the summary of these 2 models.

Table 4.2.2 Summary of model selection of full 2 for barley

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.2.2)	14	17	5.247	468.00	112.15	132.23	0.7820
(4.2.3)	7	24	5.369	691.89	110.27	120.31	0.6778

By comparing model (4.2.2) and model (4.2.3), we find model (4.2.3) has smaller AIC and BIC. Further, RMSE for the two models are close to each other and the number of parameters for model (4.2.3) is less than that of model (4.2.2). All the information tends to tell us that model (4.2.3) is better than model (4.2.2). To be prudent, we perform the F-test of reduction to check whether model (4.2.2) is able to reduce to model (4.2.3) or not and the results are given below in Table 4.2.3.

Table 4.2.3 F-test of reduction for Model (4.2.2) and Model (4.2.3)

Model	Res. Df	RSS	Df Diff	Sum of Sq	F Statistic	P value
(4.2.3)	24	691.89	-	-	-	-
(4.2.2)	17	468.00	7	223.88	1.1618	0.3736

Our p-value is 0.3736 which means we fail to reject null hypothesis (H_0: It is permissible to reduce complex model to simpler model.) and it is permissible to reduce model (4.2.2) to model (4.2.3). Before we make a conclusion that model (4.2.3) is the best model for Full 2, we would like to check whether the interaction term is necessary or not. Therefore, we have the following model (4.2.4) which is without interaction term.

$$yield = 134.5864 - 4.9352 * \text{meanT in Jul} - 1.0844 * \text{meanT in May}$$

$$+1.0731 * \text{diffT in May} + 2.4283 * \text{ppt in May} - 2.1844 * \text{ppt in Aug} \quad \textbf{(4.2.4)}$$

We perform the F test of reduction for reducing model (4.2.3) to model (4.2.4) and the results are given below in Table 4.2.4.

Table 4.2.4 F-test of reduction for Model (4.2.3) and Model (4.2.4)

Model	Res. Df	RSS	Df Diff	Sum of Sq	F Statistic	P value
(4.2.4)	25	761.00	-	-	-	-
(4.2.3)	24	691.89	1	69.115	2.3974	0.1346

Our p-value is 0.1346 which means we fail to reject null hypothesis and it is permissible to reduce model (4.2.3) to model (4.2.4), indicating that the interaction term is not necessary. Finally, we choose model (4.2.4) as our best model for Full 2. The two best models from Full 1 and Full 2 are summarized in Table 4.2.5.

Table 4.2.5 Summary of Two Best Models for Barley

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.2.1)	3	28	5.935	986.36	113.26	**117.56**	0.5406
(4.2.4)	6	25	**5.517**	**761.00**	**111.22**	119.82	**0.6456**

*Bolded value means better

By comparing model (4.2.1) and model (4.2.4), we find model (4.2.4) has both smaller AIC and RMSE although BIC is bigger than model (4.2.1). R square value for model (4.2.4) is bigger than model (4.2.1). Therefore, we select model (4.2.4) as our final best model. To be prudent, we perform the F-test of reduction to help us make the final decision in order to check whether model (4.2.4) is permissible to reduce to model (4.2.1) and the results are given below in Table 4.2.6.

Table 4.2.6 F-test of reduction for Model (4.2.4) and Model (4.2.1)

Model	Res. Df	RSS	Df Diff	Sum of Sq	F Statistic	P value
(4.2.1)	28	986.36	-	-	-	-
(4.2.4)	25	761.00	3	225.35	2.4677	0.0855

Our p -value is 0.0855. At α level 0.1, we have sufficient evidence to reject the null hypothesis and conclude that it is not permissible to reduce model (4.2.4) to model (4.2.1). Therefore, we are confident to conclude that model (4.2.4) is the best model for barley. Further, we check the residual plot and the normality of residuals.

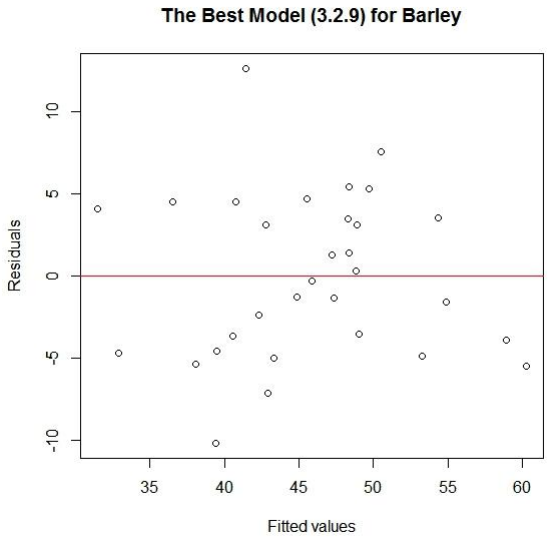


Figure 4.2.1 Residual Plot for Barley

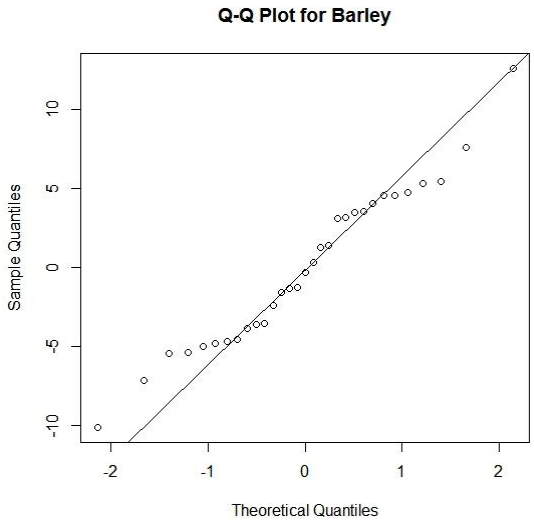


Figure 4.2.2 Q-Q Plot for Barley

Based on the residual plot and Q-Q plot, one could see the mean of the residuals is 0 and they are roughly normally distributed, while the dots in the Q-Q plot are not off the straight line too much. Shapiro-Wilk normality test is also used to check the normality of the residual. The result is shown in Table 4.2.7.

Table 4.2.7 Shapiro-Wilk Normality Test for Barley

Model	Statistics	p -value
(4.2.4)	0.9694	0.5028

Our p-value for Shapiro-Wilk normality test is 0.5028. At α level 0.1, we fail to reject the null hypothesis (H_0: Residuals have a normal distribution) and conclude that the residuals for model (4.2.4) have a normal distribution. We expect the residuals for a good model are normally distributed. The results confirm that model (4.2.4) is a good model.

We would like to interpret the model (4.2.4) in the following way. The negative coefficient for mean temperature in May and July tells us that if mean temperature in May and July is high, it will decrease barley yield. The positive coefficient for difference temperature in May tells us that if the volatility of temperature is big in May, it will benefit barley growing and will increase barley yield. We also find that the coefficient is positive for precipitation in May while it is negative for precipitation in August. That is because barley is grass-like in May and needs more water to grow up and head well. If precipitation in May is sufficient, it will benefit barley's growing and result in a larger yield. However, barley has already grown up and is ready to be harvested in August. If the precipitation in August is too much, it causes barley to rot and of course, will decrease barley yield.

4.3 CANOLA

From table 3.3.1, we find that p-values of either distance correlation test or Pearson correlation test for difference temperature in June and July, mean temperature in June and July and precipitation in May are smaller than 0.1. This indicates that all these five variables diffT in June, diffT in July, meanT in June, meanT in July and ppt in May are significant variables. Based on the definition for Full 1in the introduction part, we include these significant variables as well as the cumulative precipitation in February and March. The **Full 1** is.

$$yield = \widehat{\beta_0} + \widehat{\beta_1} * diffT\ in\ Jun + \widehat{\beta_2} * diffT\ in\ Jul + \widehat{\beta_3} * meanT\ in\ Jun$$

$$+\widehat{\beta_4} * meanT\ in\ Jul + \widehat{\beta_5} * ppt\ in\ May + \widehat{\beta_6} * ppt\ in\ Feb\ and\ Mar$$

Also, based on the definition for Full 2, we include significant variables and interaction terms as well as the cumulative precipitation in February and March. The **Full 2** is.

$$yield = \widehat{\beta_0} + \widehat{\beta_1} * diffT\ in\ May + \widehat{\beta_2} * diffT\ in\ Jun + \widehat{\beta_3} * diffT\ in\ Jul$$

$$+\widehat{\beta_4} * meanT\ in\ May + \widehat{\beta_5} * meanT\ in\ Jun + \widehat{\beta_6} * meanT\ in\ Jul + \widehat{\beta_7} * ppt\ in\ May$$

$$+\widehat{\beta_8} * ppt\ in\ Jun + \widehat{\beta_9} * ppt\ in\ Jul + \widehat{\beta_{10}} * ppt\ in\ Feb\ and\ Mar$$

$$+\widehat{\beta_{11}} * diffT\ in\ May * ppt\ in\ May + \widehat{\beta_{12}} * diffT\ in\ Jun * ppt\ in\ Jun$$

$$+\widehat{\beta_{13}} * diffT\ in\ Jul * ppt\ in\ July + \widehat{\beta_{14}} * meanT\ in\ May * ppt\ in\ May$$

$$+\widehat{\beta_{15}} * meanT\ in\ Jun * ppt\ in\ Jun + \widehat{\beta_{16}} * meanT\ in\ Jul * ppt\ in\ Jul$$

For Full 1, we use **backward selection** method first, which returns us the following model.

$$yield = 80.261 - 1.228 * \text{diffT in Jun} - 2.248 * \text{meanT in Jul} \qquad (4.3.1)$$

Then, we use **forward selection**, **stepwise selection** starts from **full** and **stepwise selection** starts from **null**. All the 4 methods return us the same best model for Full 1. Table 4.3.1 is the summary of this model.

Table 4.3.1 Summary of Model Selection of Full 1 for Canola

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.3.1)	3	24	4.084	400.25	78.8	82.69	0.4658

For full 2, we use **backward selection** method first. This returns us the following model.

$$yield = 97.1720 - 0.032 * \text{diffT in Jul} - 3.77589 * \text{meanT in Jul} - 2.47821 * \text{ppt in Jul}$$

$$-1.2177 * \text{diffT in Jul} * \text{ppt in Jul} + 0.9048 * \text{meanT in Jul} * \text{ppt in Jul} \qquad (4.3.2)$$

which is the same model we get using stepwise selection starting from the full model.

Then, we use **forward selection** method. This returns us the following model.

$$yield = 80.261 - 1.228 * \text{diffT in Jun} - 2.248 * \text{meanT in Jul} \qquad (4.3.3)$$

which is the same model we get using stepwise selection starting from the null model.

We have two options from full 2 and we need to determine which one is better. Table 4.3.2 is the summary of the two models.

Table 4.3.2 Summary of Model Selection of Full 2 for Canola

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.3.2)	6	21	4.032	341.37	80.5	88.28	0.5444
(4.3.3)	3	24	4.084	400.25	78.8	82.69	0.4658

By comparing model (4.3.2) and model (4.3.3), we find model (4.3.3) has both smaller AIC and BIC. Further, RMSE for the two models are close to each other and the number of parameters for model (4.3.3) is less than that of model (4.3.2). Therefore, we conclude that model (4.3.3) outperforms model (4.3.2). We select model (4.3.3) as our best model from Full 2.

Since model (4.3.1) from Full 1 is the same as model (4.3.3) from Full 2, it seems that model (4.3.1) might be our final best model. However, before we make that decision, let take a look at the following model (4.3.4):

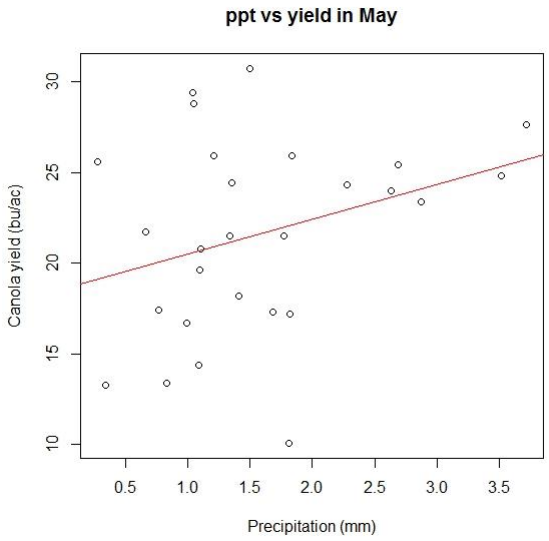
$$yield = 75.66 - 1.134 * \text{diffT in Jun} - 2.142 * \text{meanT in Jul} + 0.8461 * \text{ppt in May} \quad \textbf{(4.3.4)}$$

The difference between model (4.3.4) and model (4.3.1) is that we have one extra term ppt in May in model (4.3.4) and also, model (4.3.4) is the second best model from both Full 1 and Full 2 by using forward selection method. The summary of model (4.3.4) and model (4.3.1) is shown in Table 4.3.3.

Table 4.3.3 Summary of Model (4.3.1) and Model (4.3.4) for Canola

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.3.4)	4	23	4.101	386.84	79.9	85.06	0.4837
(4.3.1)	3	24	4.084	400.25	78.8	82.69	0.4658

From the Table 4.3.3, based on the criteria we use (i.e. AIC or BIC etc.), model selections do give us the best option which is model (4.3.1). However, we would like to argue that model (4.3.4) is our final best model based on the 3 following reasons. First, although AIC and BIC of model (4.3.4) is bigger, the RMSE of model (4.3.4) is smaller and R square of model (4.3.4) is bigger which indicate that model (4.3.4) might be a good model as well. Second, recall that p-values of both distance correlation test and Pearson correlation test for precipitation in May are small. This indicates that ppt in May might be a significant factor to impact canola yield. Lastly, let's take a look at Figure 3.3.5 again.



From Figure 3.3.5, we find there is a very clear upward trend which implies ppt in May positively impacts canola yield and model (4.3.4) reveals this fact perfectly since the coefficient of ppt in May is positive! However, model (4.3.1) fails to tell us that! Therefore, we are confident to conclude that model (4.3.4) is the best model for canola. Further, we check the residual plot and the normality of residuals.

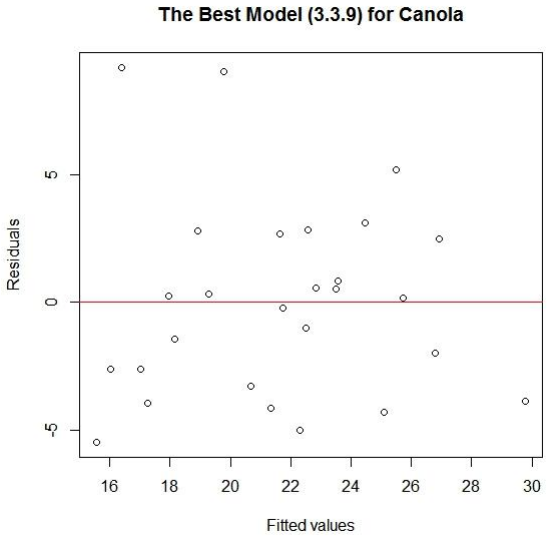


Figure 4.3.1 Residual Plot for Canola

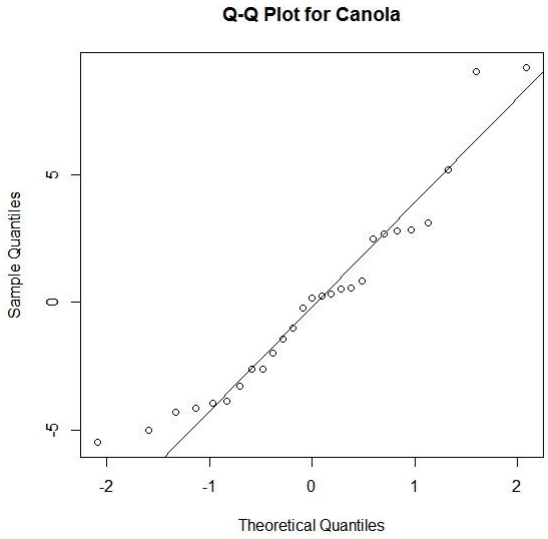


Figure 4.3.2 Q-Q Plot for Canola

Based on the residual plot and Q-Q plot, we would say that the residuals may not have a good normal distribution. The mean of the residuals is a little bit off 0 while the dots in the Q-Q plot are off the straight line as well. Shapiro-Wilk normality test is also used to check the normality of the residual. The result is shown in the Table 4.3.4.

Table 4.3.4 Shapiro-Wilk Normality Test for Canola

Model	Statistics	p -value
(4.3.4)	0.9318	0.0766

Our p -value for Shapiro-Wilk normality test is 0.0766. At α level 0.1 (it really depends on α level), we reject the null hypothesis and conclude that (H_a: Residuals do not have a normal distribution) the residuals for model (4.3.4) do not have a normal distribution. We expect the residuals for a good model to have a normal distribution but still, we choose model (4.3.4) as our best model for canola.

We would like to interpret the model (4.3.4) in the following ways. The negative coefficient for difference temperature in June tells us that if the volatility of temperature is big in June, it will not benefit canola growing and will decrease canola yield. The negative coefficient for mean temperature in July tells us that if the mean temperature in July is high, it will not benefit canola growing and will also decrease canola yield. We find that coefficient is positive for precipitation in May. Our explanation is that canola in May needs more water to grow up. If precipitation in May is sufficient, it will benefit canola growing and result in large yield.

4.4 Flax

From Table 3.4.1, we find that p-values of both distance correlation test and Pearson correlation test for difference temperature in May, June and July, mean temperature in July and precipitation in June are smaller than 0.1. This indicates that all these five variables diffT in May, diffT in June, diffT in July, meanT in July and ppt in June are significant variables. Based on the definition for Full 1in the introduction part, we include these significant variables as well as the cumulative precipitation in February and March. The **Full 1** is.

$$yield = \widehat{\beta_0} + \widehat{\beta_1} * diffT\ in\ May + \widehat{\beta_2} * diffT\ in\ Jun + \widehat{\beta_3} * diffT\ in\ Jul$$

$$+\widehat{\beta_4} * meanT\ in\ Jul + \widehat{\beta_5} * ppt\ in\ Jun + \widehat{\beta_6} * ppt\ in\ Feb\ and\ Mar$$

Also, based on the definition for Full 2, we include significant variables and interaction terms as well as the cumulative precipitation in February and March. The **Full 2** is.

$$yield = \widehat{\beta_0} + \widehat{\beta_1} * diffT\ in\ May + \widehat{\beta_2} * diffT\ in\ Jun + \widehat{\beta_3} * diffT\ in\ Jul$$

$$+\widehat{\beta_4} * meanT\ in\ Jun + \widehat{\beta_5} * meanT\ in\ Jul + \widehat{\beta_6} * ppt\ in\ May + \widehat{\beta_7} * ppt\ in\ Jun$$

$$+\widehat{\beta_8} * ppt\ in\ Jul + \widehat{\beta_9} * ppt\ in\ Feb\ and\ Mar + \widehat{\beta_{10}} * diffT\ in\ May * ppt\ in\ May$$

$$+\widehat{\beta_{11}} * diffT\ in\ Jun * ppt\ in\ Jun + \widehat{\beta_{12}} * diffT\ in\ Jul * ppt\ in\ Jul$$

$$+\widehat{\beta_{13}} * meanT\ in\ Jun * ppt\ in\ Juln + \widehat{\beta_{14}} * meanT\ in\ Jul * ppt\ in\ Jul$$

For Full 1, we use **backward selection** method first. This returns us the following model.

$$yield = 55.067 - 0.896 * \text{diffT in May} - 1.447 * \text{meanT in Jul} + 1.542 * \text{ppt in Jun} \quad \textbf{(4.4.1)}$$

which is the same model we get using stepwise selection starting from full model.

Then, we use **forward selection** method. This returns us the following model.

$$yield = 61.066 - 1.17 * \text{diffT in Jul} - 1.157 * \text{diffT in Jun} - 0.75 * \text{diffT in May} \quad (4.4.2)$$

which is the same model we get using stepwise selection starting from null model.

We have two options from Full 1 and we need to determine which one is better. Table 4.4.1 is the summary of the 2 models.

Table 4.4.1 Summary of Model Selection of Full 1 for Flax

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.4.1)	4	27	3.385	309.43	79.323	85.059	0.4998
(4.4.2)	4	27	3.635	356.72	83.732	89.468	0.4233

By comparing model (4.4.1) and model (4.4.2), we find model (4.4.1) has smaller AIC, BIC and RMSE as well as RSS. R square for model (4.4.1) is also bigger than model (4.4.2). Further, the two models have the same parameters and residual degree freedom. All the information indicates that model (4.4.1) perfectly outperform model (4.4.2). Therefore, we select model (4.4.1) as our best model from Full 1.

For Full 2, we use **backward selection** method first. This returns us the following model.

$$yield = 80.903 - 1.1 * \text{diffT in Jul} - 2.437 * \text{diffT in Jun} - 0.667 * \text{diffT in May}$$

$$-0.0186 * \text{meanT in Jul} - 0.733 * \text{meanT in Jun} - 17.823 * \text{ppt in Jun}$$

$$+1.968 * \text{ppt in Jul} + 1.04136 * \text{diffT in Jul} * \text{ppt in Jul} + 0.793 * \text{diffT in Jun} * \text{ppt in Jun}$$

$$-0.766 * \text{meanT in Jul} * \text{ppt in Jul} + 0.577 * \text{meanT in Jun} * \text{ppt in Jun} \quad (4.4.3)$$

which is the same model we get using stepwise selection starting from full model.

Then, we use **forward selection** method. This returns us the following model.

$$yield = 56.447 - 1.26 * \text{diffT in Jul} - 1.454 * \text{diffT in Jun} - 0.754 * \text{diffT in May}$$

$$+0.61 * \text{meanT in Jun} \quad (4.4.4)$$

which is the same model we get using stepwise selection starting from null model.

We have two options from Full 2 and we need to determine which one is better. Table 4.4.2 is the summary of the two models.

Table 4.4.2 Summary of Model Selection of Full 2 for Flax

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.4.3)	12	19	3.434	224.11	85.322	102.53	0.6377
(4.4.4)	5	26	3.538	325.50	82.893	90.063	0.4738

By comparing model (4.4.3) and model (4.4.4), we find model (4.4.4) has both smaller AIC and BIC. Further, RMSE for the two models are close to each other and the number of parameters for model (4.4.4) is much less than model (4.4.3). These information tells us that

model (4.4.4) is better than model (4.4.3). To be prudent, we perform F-test of reduction to check whether model (4.4.3) is permissible to reduce to model (4.4.4) or not and the results are given below in Table 4.4.3.

Table 4.4.3 F-test of Reduction for Model (4.4.5) and Model (4.4.6)

Model	Res. Df	RSS	Df Diff	Sum of Sq	F Statistic	P value
(4.4.4)	26	325.50	-	-	-	-
(4.4.3)	19	224.11	7	101.39	1.228	0.3359

Our p -value is 0.3359 which means we fail to reject null hypothesis (H_0: It is permissible to reduce complex model to simpler model.) and it is permissible to reduce model (4.4.3) to model (4.4.4). Therefore, we conclude that model (4.4.4) outperforms model (4.4.3) and we select model (4.4.4) as our best model from Full 2. The two best models from Full 1 and Full 2 are summarized in Table 4.4.4.

Table 4.4.4 Summary of Two Best Models for Flax

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.4.1)	4	27	3.385	309.43	79.323	85.059	0.4998
(4.4.4)	5	26	3.538	325.5	82.893	90.063	0.4738

By comparing model (4.4.1) and model (4.4.4), we find model (4.4.1) has smaller AIC, BIC and RMSE as well as RSS. R square for model (4.4.1) is also bigger than model (4.4.4). Further, model (4.4.1) has less parameter than model (4.4.4). All the information indicates that model (4.4.1) perfectly outperform model (4.4.4). Therefore, we select model (4.4.1) as our final best model. Further, we check the residual plot and the normality of residuals.

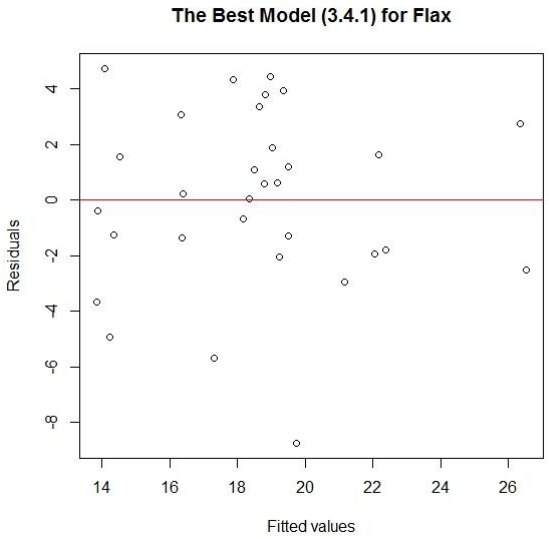


Figure 4.4.1 Residual Plot for Flax

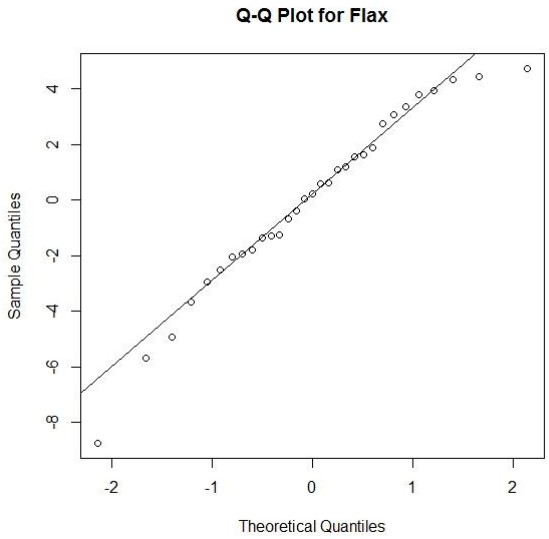


Figure 4.4.2 Q-Q Plot for Flax

Based on the residual plot and Q-Q plot, one could see the mean of the residuals is roughly 0 and also, the residuals are normal distributed while the dots in the Q-Q plot match the straight line very well. Shapiro-Wilk normality test is also used to check the normality of the residual. The result is shown in the Table 4.4.5.

Table 4.4.5 Shapiro-Wilk Normality Test for Flax

Model	Statistics	p -value
(4.4.1)	0.9616	0.3223

Our p -value for Shapiro-Wilk normality test is 0.3223. At α level 0.1, we have insufficient evidence to reject the null hypothesis and conclude that the residuals for model (4.4.1) have a normal distribution. We expect the residuals for a good model to have a normal distribution. Therefore, based on the previous analysis, we are confident to choose model (4.4.1) as our best model for flax.

Here are our explanations for model (4.4.1). The negative coefficient for difference temperature in May tells us that if the volatility of temperature is big in May, it will not benefit flax growing and will decrease flax yield. The negative coefficient for mean temperature in July tells us that if the mean temperature in July is high, it will not benefit flax growing and will decrease flax yield as well. We also find that coefficient is positive for precipitation in June. Our explanation is that flax in June needs more water to grow up. If precipitation in June is sufficient, it will benefit flax growing and result in greater yield.

4.5 Oats

From Table 3.5.1, we find that p -values of either distance correlation test or Pearson correlation test for difference temperature in June and July, mean temperature in July and precipitation in May are smaller than 0.1. This indicates that all these four variables diffT in June, diffT in July, meanT in Jul and ppt in May are significant variables. Based on the definition for Full , we include these significant variables as well as the cumulative precipitation in February and March. The **Full 1** is.

$$yield = \widehat{\beta_0} + \widehat{\beta_1} * diffT\ in\ Jun + \widehat{\beta_2} * diffT\ in\ Jul + \widehat{\beta_3} * meanT\ in\ Jul$$

$$+\widehat{\beta_4} * ppt\ in\ May + \widehat{\beta_5} * ppt\ in\ Feb\ and\ Mar$$

Also, based on the definition for Full 2, we include significant variables and interaction terms as well as the cumulative precipitation in February and March. The **Full 2** is.

$$yield = \widehat{\beta_0} + \widehat{\beta_1} * diffT\ in\ May + \widehat{\beta_2} * diffT\ in\ Jun + \widehat{\beta_3} * diffT\ in\ Jul$$

$$+\widehat{\beta_4} * meanT\ in\ May + \widehat{\beta_5} * meanT\ in\ Jul + \widehat{\beta_6} * ppt\ in\ May + \widehat{\beta_7} * ppt\ in\ Jun$$

$$+\widehat{\beta_8} * ppt\ in\ Jul + \widehat{\beta_9} * ppt\ in\ Feb\ and\ Mar + \widehat{\beta_{10}} * diffT\ in\ May * ppt\ in\ May$$

$$+\widehat{\beta_{11}} * diffT\ in\ Jun * ppt\ in\ Jun + \widehat{\beta_{12}} * diffT\ in\ Jul * ppt\ in\ July$$

$$+\widehat{\beta_{13}} * meanT\ in\ May * ppt\ in\ May + \widehat{\beta_{14}} * meanT\ in\ Jul * ppt\ in\ Jul$$

For Full 1, we use **backward selection** method first, which returns.

$$yield = 166.1 - 2.514 * \text{diffT in July} - 2.148 * \text{diffT in Jun} - 2.653 * \text{meanT in Jul} \quad (4.5.1)$$

Then, we use **forward selection**, **stepwise selection** starts from **full** and **stepwise selection** starts from **null**. All the 4 methods return us the same best model for Full 1. Table 4.5.1 is the summary of this model.

Table 4.5.1 Summary of Model Selection of Full 1 for Oats

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.5.1)	4	27	7.6767	1590.8	130.08	135.814	0.5015

For Full 2, we use **backward selection** method first. This returns us the following model.

$$yield = 178.494 - 2.142 * \text{diffT in Jun} - 3.617 * \text{meanT in Jul} - 2.87 * \text{meanT in May}$$
$$-10.826 * \text{ppt in May} + 1.144 * \text{meanT in May} * \text{ppt in May} \quad (4.5.2)$$

which is the same model we get using stepwise selection starting from full model.

Then, we use **forward selection** method. This returns us the following model.

$$yield = 166.1 - 2.514 * \text{diffT in July} - 2.148 * \text{diffT in Jun} - 2.653 * \text{meanT in Jul} \quad (4.5.3)$$

which is the same model we get using stepwise selection starting from null model.

We have two options from Full 2 and we need to determine which one is better. Table 4.5.2 is the summary of the two models.

Table 4.5.2 Summary of Model Selection of Full 2 for Oats

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.5.2)	6	25	7.394	1366.8	129.37	137.98	0.5718
(4.5.3)	4	27	7.676	1590.8	130.08	135.81	0.5015

By comparing model (4.5.2) and model (4.5.3), we find model (4.5.3) has smaller BIC while AIC is a little bit bigger than model (4.5.2). Further, RMSE for the two models are close to each other. Generally speaking, the performance for model (4.5.2) and model (4.5.3) are comparable. Since the number of parameters for model (4.5.3) is 2 less than model (4.5.2), we tend to choose model (4.5.3) as our best model from Full 2.

The best model (4.5.1) from Full 1 is the same as the best model (4.5.3) from Full 2. It seems that we may conclude model (4.5.1) as our final best model. However, before we make that decision, let take a look at the following model (4.5.4):

$$yield = 155.85 - 1.859 * \text{diffT in Jul} - 2.323 * \text{diffT in Jun} - 2.645 * \text{meanT in Jul}$$

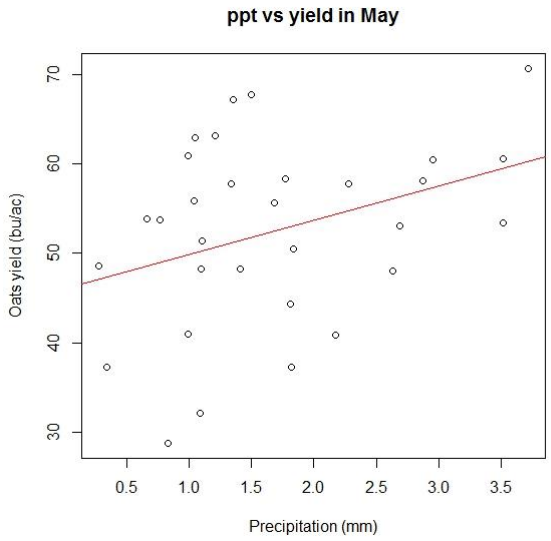
$$+1.992 * \text{ppt in May} \qquad \textbf{(4.5.4)}$$

The difference between model (4.5.4) and model (4.5.1) is that we have one extra term ppt in May in model (4.5.4) and indeed, model (4.5.4) is the second best model from both Full 1 and Full 2 by using forward selection method. The summary of model (4.5.4) and model (4.5.1) is shown in Table 4.5.3.

Table 4.5.3 Summary of Model (4.5.1) and Model (4.5.9) for Oats

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.5.4)	5	26	7.609	1505.5	130.37	137.54	0.5283
(4.5.1)	4	27	7.676	1590.8	130.08	135.81	0.5015

From the Table 4.5.3, based on the criteria we use (i.e. AIC or BIC etc.), model selections do give us the best option which is model (4.5.1). However, we would like to argue that model (4.5.4) is our final best model based on the 3 reasons. First, although AIC and BIC of model (4.5.4) is bigger, the RMSE of model (4.5.4) is smaller and R square of model (4.5.4) is bigger. All these indicate that model (4.5.4) might be a good model as well. Second, recall that p -values of Pearson correlation test for precipitation is 0.057. This indicates that ppt in May is a significant factor to impact oats yield and we should include it. Lastly, let's take a look at Figure 3.5.5 again.



From Figure 3.5.5, we find there is a very clear upward trend which implies ppt in May positively impacts oats yield and model (4.5.4) reveals this fact perfectly since the coefficient of ppt in May is positive (1.992). However, the model (4.5.1) fails to tell us that! Therefore, we conclude that model (4.5.4) is the final best model for oats. Further, we check the residual plot and the normality of residuals.

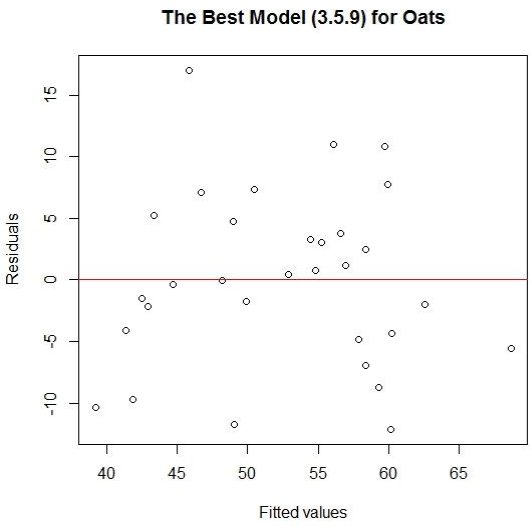


Figure 4.5.1 Residual Plot for Oats

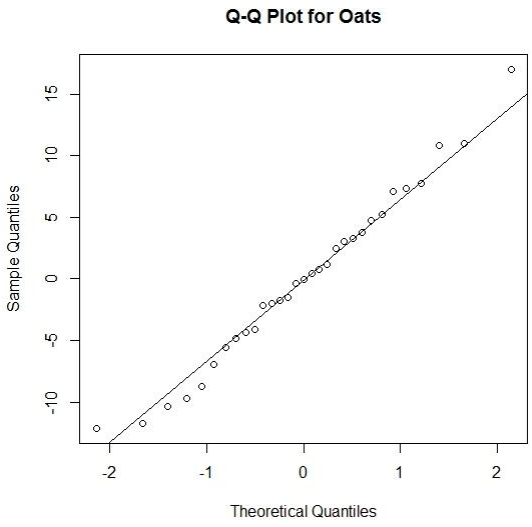


Figure 4.5.2 Q-Q Plot for Oats

Based on the residual plot and Q-Q plot, one would say that the residuals have a very good normal distribution. The mean of the residuals is roughly 0 and the dots in the Q-Q plot match the straight line very well. Shapiro-Wilk normality test is also used to check the normality of the residual. The result is shown in the Table 4.5.4 as below.

Table 4.5.4 Shapiro-Wilk Normality Test for Oats

Model	Statistics	p -value
(4.5.4)	0.9834	0.8987

Our p -value for Shapiro-Wilk normality test is 0.8987. At α level 0.1, we have insufficient evidence to reject the null hypothesis and conclude the residuals for model (4.5.4) have a normally distribution. We expect the residuals for a good model to have a normal distribution. Therefore, the model (4.5.4) is our final best model to predict oats yield.

In the model (4.5.4), the coefficients for difference temperature in both June and July are negative which imply that if the volatility of temperature is big in June or July, it will not benefit oats growing and will decrease oats yield. The negative coefficient for mean temperature in July tells us that if the mean temperature in July is high, it will not benefit oats growing and will also decrease oats yield. Finally, we have positive a coefficient for precipitation in May. Our explanation is that oats in May needs more water to grow up. If precipitation in May is sufficient, it will benefit oats growing and result in large yield.

4.6 Pea

From Table 3.6.1, we find that p-values of either distance correlation test or Pearson correlation test for difference temperature in June, mean temperature in May and June and precipitation in April, May and June are smaller than 0.1. This indicates that all these six variables diffT in June, meanT in May, meanT in Jun, ppt in April, ppt in May and ppt in June are significant variables. Based on the definition for Full 1in the introduction part, we include these significant variables as well as the cumulative precipitation in February and March.

The **Full 1** is.

$$yield = \widehat{\beta_0} + \widehat{\beta_1} * diffT\ in\ Jun + \widehat{\beta_2} * meanT\ in\ May + \widehat{\beta_3} * meanT\ in\ Jun$$

$$+\widehat{\beta_4} * ppt\ in\ Apr + \widehat{\beta_5} * ppt\ in\ May + \widehat{\beta_6} * ppt\ in\ Jun + \widehat{\beta_7} * ppt\ in\ Feb\ and\ Mar$$

Also, based on the definition for Full 2, we include significant variables and interaction terms as well as the cumulative precipitation in February and March. The **Full 2** is.

$$yield = \widehat{\beta_0} + \widehat{\beta_1} * diffT\ in\ Apr + \widehat{\beta_2} * diffT\ in\ May + \widehat{\beta_3} * diffT\ in\ Jun$$

$$+\widehat{\beta_4} * meanT\ in\ Apr + \widehat{\beta_5} * meanT\ in\ May + \widehat{\beta_6} * meanT\ in\ Jun + \widehat{\beta_7} * ppt\ in\ Apr$$

$$+\widehat{\beta_8} * ppt\ in\ May + \widehat{\beta_9} * ppt\ in\ Jun + \widehat{\beta_{10}} * ppt\ in\ Feb\ and\ Mar$$

$$+\widehat{\beta_{11}} * diffT\ in\ Apr * ppt\ in\ Apr + \widehat{\beta_{12}} * diffT\ in\ May * ppt\ in\ May$$

$$+\widehat{\beta_{13}} * diffT\ in\ Jun * ppt\ in\ Jun + \widehat{\beta_{14}} * meanT\ in\ Apr * ppt\ in\ Apr$$

$$+\widehat{\beta_{15}} * meanT\ in\ May * ppt\ in\ May + \widehat{\beta_{16}} * meanT\ in\ Jun * ppt\ in\ Jun$$

For Full 1, we use **backward selection** method first. This returns us the following model.

$$yield = 88.66 + 4.364 * \text{ppt in Feb and Mar} + 3.657 * \text{ppt in Apr} - 1.991 * \text{diffT in Jun}$$

$$-1.109 * \text{meanT in May} - 1.869 * \text{meanT in Jun} \quad \textbf{(4.6.1)}$$

Then, we use **forward selection** method, **stepwise selection** method starts from **full** and **stepwise selection** method starts from **null**. All the 4 methods return us the same best model for Full 1. Table 4.6.1 is the summary of model (4.6.1).

Table 4.6.1 Summary of Model Selection of Full 1 for Pea

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.6.1)	6	14	4.032	227.60	60.637	66.612	0.8252

For Full 2, we use **backward selection** method first. This returns us the following model.

$$yield = 30.178 + 4.0812 * \text{ppt in Feb and Mar} - 3.611 * \text{diffT in Apr}$$

$$-3.577 * \text{diffT in May} + 0.3428 * \text{meanT in May} - 2.077 * \text{meanT in Jun}$$

$$+2.727 * \text{meanT in Apr} - 24.194 * \text{ppt in Apr} + 47.281 * \text{ppt in May}$$

$$+1.846 * \text{ppt in Jun} + 3 * \text{diffT in Apt} * \text{ppt in Apr} - 2.423 * \text{diffT in May} * \text{ppt in May}$$

$$-1.149 * \text{meanT in May} * \text{ppt in May} - 2.549 * \text{meanT in Apr} * \text{ppt in Apr} \quad (4.6.2)$$

which is the same model we get using stepwise selection starting from full model.

Then, we use **forward selection** method. This returns us the following model.

$$yield = 83.866 - 2.222 * \text{meanT in Jun} - 0.9278 * \text{diffT in May} - 1.381 * \text{meanT in May}$$

$$+1.141 * \text{ppt in Jun} + 2.548 * \text{ppt in Apr} + 3.522 * \text{ppt in Feb and Mar} \quad (4.6.3)$$

which is the same model we get using stepwise selection starting from null model.

We have two options from full 2 and we need to determine which one is better. Table 4.6.2 is the summary of the two models.

Table 4.6.2 Summary of Model Selection of Full 2 for Pea

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.6.2)	14	6	4.915	144.94	67.612	81.552	0.8887
(4.6.3)	7	13	4.22	231.54	62.98	69.95	0.8222

By comparing model (4.6.2) and model (4.6.3), we find model (4.6.3) has both smaller AIC and BIC. RMSE for model (4.6.3) is also smaller than model (4.6.2). Further, the number of parameters for model (4.6.3) is less than model (4.6.2). All the information tells us that model (4.6.3) perfectly outperforms model (4.6.2). Therefore, we choose model (4.6.3) as our best model from full2. The two best models from full 1 and full 2 are summarized in Table 4.6.3.

Table 4.6.3 Summary of Two Best Models for Pea

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.6.1)	6	14	4.032	227.60	60.637	66.612	0.8252
(4.6.3)	7	13	4.22	231.54	62.980	69.950	0.8222

By comparing model (4.6.1) and model (4.6.3), we find model (4.6.1) has smaller AIC, BIC and RMSE as well as RSS. R square for model (4.6.1) is also bigger than model (4.6.3). Further, model (4.6.1) has less parameter than model (4.6.3). All the information indicates that model (4.6.1) perfectly outperform model (4.6.3). Therefore, we select model (4.6.1) as our final best model. Further, we check the residual plot and the normality of residuals.

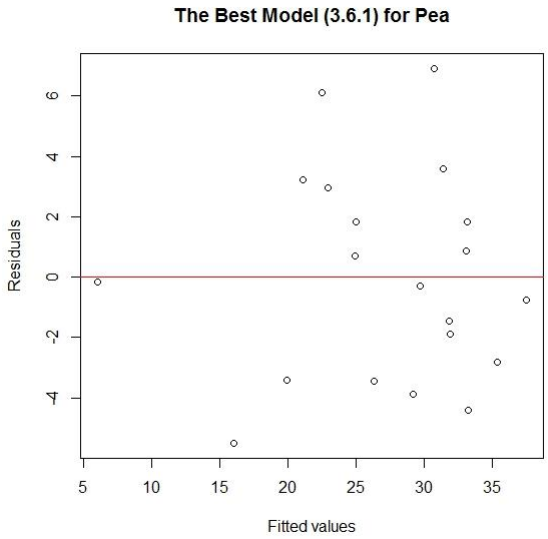


Figure 4.6.1 Residual Plot for Pea

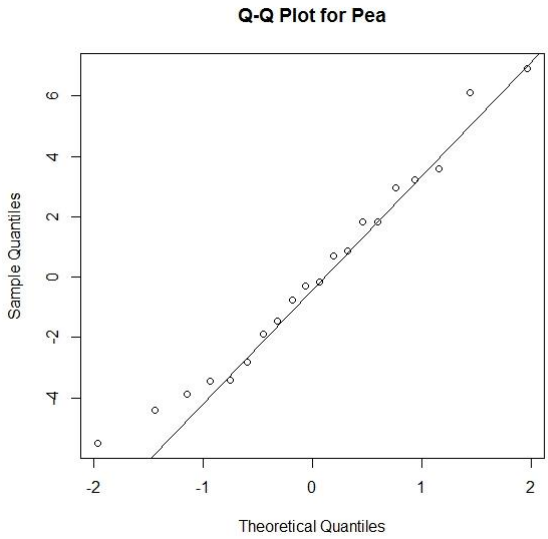


Figure 4.6.2 Q-Q Plot for Pea

Based on the residual plot and Q-Q plot, one could see the mean of the residuals is roughly 0 and also and the residuals are normally distributed while the dots in the Q-Q plot match the straight line well. Shapiro-Wilk normality test is also used to check the normality of the residual. The result is shown in the Table 4.6.4.

Table 4.6.4 Shapiro-Wilk Normality Test for Pea

Model	Statistics	*p* -value
(4.6.1)	0.9697	0.7493

Our p -value for Shapiro-Wilk normality test is 0.7493. At α level 0.1, we have insufficient evidence to reject the null hypothesis and conclude that the residuals for model (4.6.1) have a normal distribution. We expect the residuals for a good model to have a normal distribution. Therefore, based on the previous analysis, we are confident to choose model (4.6.1) as our best model for pea and it will predict pea yield well.

The model (4.6.1) is quite different from the previous ones for other crops. We find that the cumulative precipitation in February and March is included in the model and the coefficient for which is positive. This tells us, before peas are seeded, if there is sufficient water in the soil, it will benefit pea growing and result in greater yield! The coefficient is positive for precipitation in April as well which means that peas in April need more water to grow up. The negative coefficient for difference temperature in June tells us that if the volatility of temperature is big in June, it will not benefit pea growing and will decrease pea yield. The negative coefficient for mean temperature in both May and June tells us that if the mean temperature in May or June is high, it will not benefit fpea growing and will result in reducing pea yield.

4.7 Spring Wheat

From Table 3.7.1, we find that p-values of either distance correlation test or Pearson correlation test for difference temperature in June and July, mean temperature in June and July and precipitation in May are smaller than 0.1. This indicates that all these five variables diffT in Jun, diffT in Jul, meanT in Jun, meanT in Jul and ppt in May are significant variables. Based on the definition for Full 1in the introduction part, we include these significant variables as well as the cumulative precipitation in February and March. The **Full 1** is defined as below.

$$yield = \widehat{\beta_0} + \widehat{\beta_1} * diffT\ in\ Jun + \widehat{\beta_2} * diffT\ in\ Jul + \widehat{\beta_3} * meanT\ in\ Jun$$

$$+\widehat{\beta_4} * meanT\ in\ Jul + \widehat{\beta_5} * ppt\ in\ May + \widehat{\beta_6} * ppt\ in\ Feb\ and\ Mar$$

Also, based on the definition for Full 2, we include significant variables and interaction terms as well as the cumulative precipitation in February and March. The **Full 2** is.

$$yield = \widehat{\beta_0} + \widehat{\beta_1} * diffT\ in\ May + \widehat{\beta_2} * diffT\ in\ Jun + \widehat{\beta_3} * diffT\ in\ Jul$$

$$+\widehat{\beta_4} * meanT\ in\ May + \widehat{\beta_5} * meanT\ in\ Jun + \widehat{\beta_6} * meanT\ in\ Jul + \widehat{\beta_7} * ppt\ in\ May$$

$$+\widehat{\beta_8} * ppt\ in\ Jun + \widehat{\beta_9} * ppt\ in\ Jul + \widehat{\beta_{10}} * ppt\ in\ Feb\ and\ Mar$$

$$+\widehat{\beta_{11}} * diffT\ in\ May * ppt\ in\ May + \widehat{\beta_{12}} * diffT\ in\ Jun * ppt\ in\ Jun$$

$$+\widehat{\beta_{13}} * diffT\ in\ Jul * ppt\ in\ July + \widehat{\beta_{14}} * meanT\ in\ May * ppt\ in\ May$$

$$+\widehat{\beta_{15}} * meanT\ in\ Jun * ppt\ in\ Jun + \widehat{\beta_{16}} * meanT\ in\ Jul * ppt\ in\ Jul$$

For Full 1, we use **backward selection** method first. This returns us the following model.

$$yield = 66.727 - 0.8975 * \text{diffT in Jun} - 1.536 * \text{meanT in Jul} + 1.781 * \text{ppt in May} \quad \textbf{(4.7.1)}$$

Then, we use **forward selection**, **stepwise selection** starts from **full** and **stepwise selection** starts from **null**. All the 4 methods return us the same best model for Full 1. Table 4.7.1 is the summary of model (4.7.1).

Table 4.7.1 Summary of Model Selection of Full 1 for Spring Wheat

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.7.1)	4	27	3.983	428.33	89.403	95.139	0.4312

For Full 2, we use **backward selection** method first. This returns us the following model.

$$yield = 118.038 + 2.494 * \text{ppt in Feb and Mar} - 1.523 * \text{diffT in Jul} - 0.817 * \text{meanT in Jun}$$
$$-1.691 * \text{meanT in Jul} - 2.374 * \text{meanT in May} - 11.607 * \text{ppt in Jul} - 13.355 * \text{ppt in May}$$
$$+0.87 * \text{diffT in Jul} * \text{ppt in Jul} + 1.26 * \text{meanT in May} * \text{ppt in May} \quad (4.7.2)$$

which is the same model we get using stepwise selection starting from full model.

Then, we use **forward selection** method. This returns us the following model.

$$yield = 66.727 - 0.8975 * \text{diffT in Jun} - 1.536 * \text{meanT in Jul} + 1.781 * \text{ppt in May} \quad (4.7.3)$$

which is the same model we get using stepwise selection starting from null model.

We have two options from full 2 and we need to determine which one is better. Table 4.7.2 is the summary of the two models.

Table 4.7.2 Summary of Model Selection of Full 2 for Spring Wheat

Model	Parm #	df	RMSE	RSS	AIC	BIC	R-Sq
(4.7.2)	10	21	3.372	238.72	83.280	97.620	0.683
(4.7.3)	4	27	3.983	428.33	89.403	95.139	0.4312

By comparing model (4.7.2) and model (4.7.3), we find model (4.7.2) has smaller AIC while model (4.7.3) has smaller BIC. RMSE for model (4.7.2) is smaller than model (4.7.3) and R square of model (4.7.2) is bigger. However, the number of parameter of model (4.7.2) is more than model (4.7.3). It seems that we are hard to make a decision since each model has advantages. Here, we add one more term diffT in Jun in model (4.7.2) so that model (4.7.3) becomes a sub-model of model (4.7.2). Then we perform F-test of reduction to check whether it is permissible to reduce model (4.7.2) to model (4.7.3). The results are given below in Table 4.7.3.

Table 4.7.3 F-test of reduction for Model (4.7.5) and Model (4.7.6)

Model	Res. Df	RSS	Df Diff	Sum of Sq	F Statistic	P value
(4.7.3)	27	428.33	-	-	-	-
(4.7.2)	20	231.37	7	196.96	2.4323	0.0563

Our p-value is 0.0.0563. At α level 0.1, we reject null hypothesis and conclude it is not permissible to reduce model (4.7.2) to model (4.7.3). However, we argue that if we choose α level 0.05, then we will accept the null so that we will choose model (4.7.3). Further, because of the principle of parsimony which tells us things are usually connected or behave in the simplest or most economical way especially with reference to alternative. Therefore, we choose the model (4.7.3) with less number of parameters as our best model for Full 2. Since model (4.7.3) is the same as model (4.7.1), we have the same best model for Full 1 and Full 2. Therefore, we

conclude that the model (4.7.1) is our final best model and below are the residual plot and the Q-Q plot for model (4.7.1).

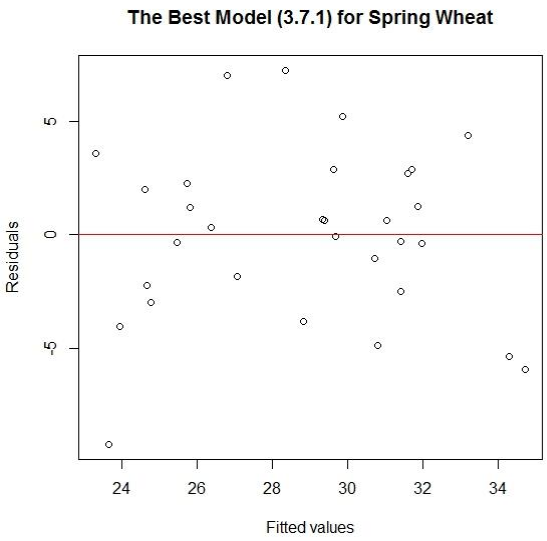


Figure 4.7.1 Residual Plot for Spring Wheat

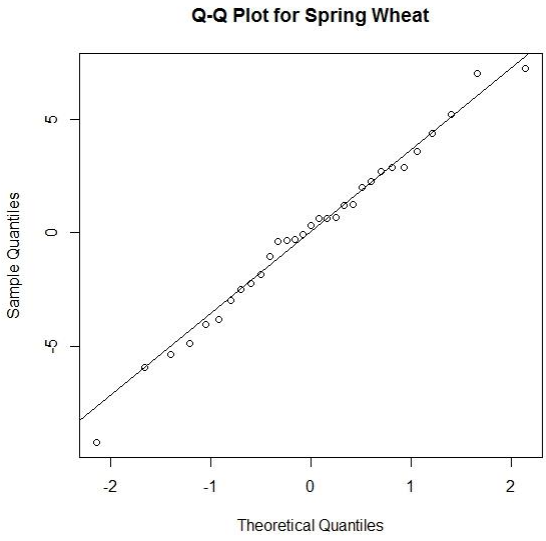


Figure 4.7.2 Q-Q Plot for Spring Wheat

Based on the residual plot and Q-Q plot, one could see the mean of the residuals is roughly 0 and also and the residuals are normally distributed while the dots in the Q-Q plot match the straight line well. Shapiro-Wilk normality test is also performed to check the normality of the residual. The result is shown in the Table 4.7.4.

Table 4.7.4 Shapiro-Wilk Normality Test for Spring Wheat

Model	Statistics	p -value
(4.7.1)	0.987	0.962

Our p -value for Shapiro-Wilk normality test is 0.962. At α level 0.1, we have insufficient evidence to reject the null hypothesis and conclude that the residuals for model (4.7.1) have a normal distribution. We expect the residuals for a good model to have a normal distribution.

Therefore, based on the previous analysis, we are confident to choose model (4.7.1) as our final best model for spring wheat and it is able to predict spring wheat yield well.

In the model (4.7.1), the negative coefficient for difference temperature in June tells us that if the volatility of temperature is big in June, it will not benefit spring wheat growing and will decrease spring wheat yield. The negative coefficient for mean temperature in July tells us that if the mean temperature in July is high, it will not benefit spring wheat growing and will also decrease spring wheat yield. We also find that coefficient is positive for precipitation in May. Our explanation is that spring wheat in May needs more water to grow up. If precipitation in May is sufficient, it will benefit spring wheat growing and result in large yield.

CHAPTER 5

CONCLUSION

In this thesis, we introduce distance correlation as a new statistical tool and compare it with Pearson correlation by simulating data for both linear and nonlinear. We use both distance correlation and Pearson correlation to help us determine two full models for each crop. Then we perform backward selection, forward selection, stepwise selection from null and stepwise selection from full on these full models. We take consideration of various factors such as RMSE, RSS, AIC, BIC, R square, number of parameters and so on and select the best model for each crop. Further, we use several statistical methods to check the goodness of fit of our best model such as Q-Q plot, residual plots etc. Finally, we have 6 best models for 6 different crops and these 6 best models are summarized in Table 5.1.

Table 5.1 Summary of Best Models

	Yield					
	Barley	Canola	Flax	Oat	Pea	Spr. Wheat
Intercept	134.586	75.660	55.067	155.850	88.660	66.727
Cum. ppt	-	-	-	-	4.364	-
Apr ppt	-	-	-	-	3.657	-
May ppt	2.428	0.846	-	1.992	-	1.781
Jun ppt	-	-	1.542	-	-1.991	-
Aug ppt	-2.184	-	-	-	-	-
May diffT	1.073	-	-0.896	-	-	-
Jun diffT	-	-1.134	-	-2.323	-	-0.898
Jul diffT	-	-	-	-1.859	-	-
May meanT	-1.084	-	-	-	-1.109	-
Jun meanT	-	-	-	-	1.869	-
Jul meanT	-4.935	-2.142	-1.447	-2.645	-	-1.536

CHAPTER 6

FURTHER DIRECTION

The study of improving crop yield is very complicated. Actually, in addition to temperature and precipitation, there are many other factors could significantly impact crop yield.

First, sunshine. What time is sunrise and what time is sunset for everyday? Even sunny or cloudy days may affect yield. Normally, longer sunshine results in greater crop yield. Second, soil potential of hydrogen. Different crops may have different favorable conditions. Normally, the yield for a certain crop would be maximized at a certain Ph. Third, fertilizer. This is not a natural factor but it still impacts crop yield. Normally, too much or too little fertilizer will negatively affect crop yield. Last, but not the least, planting density. The density should be optimized. There might be an ideal density for a crop at which the yield would be maximized. If we were supplied by these factors in addition to climate variables, we believe we would have more accurate models so that we are able to predict crop yield better.

BIBLIOGRPHY

1, Gabor J. Szekely, Maria L. Rizzo. (2007). "Measuring and testing dependence by correlation of distance" *The Annals of Statistics*. Vol 35, 2769-2794.

2, Gabor J.Szekely, Maria L. Rizzo. (2013). "The distance correlation t-test of independence in high dimension" *Journal of Multivariate Anaysis* 117, 193-213.

3, Gabor J.Szekely, Maria L. Rizzo. (2009). "Brownian Distance Covariance" *The Annals of Applied Statistics.* Vol 3, 1236-1265.

4, Michael Clark, " A comparison of correlation measures". Note.

5, David B Lobell, Chrisopher B Field. (2007). "Global scale climate crop yield ralationships and the impacts of recent warming" *Environmental Research Letter*. 2, 014002.

6, Adam Jaeger, "Response of Canadian Crop Yield to Climate Change". Note.

7, Lu Q., Lund R., Seymour L. (2005), "An update of U.S. Temperature Trends" *Journal of Climate*, 18, 4906-4914.

8, Darren Williams, "Monitoring expense report errors: control charts under independence and dependence". Note.