< 연구결과 요약 >

과 제 명	농산물 빅 데이터를 활용한 배추 가격 예측
연구목표	다양한 통계 이론을 학습한다. R프로그래밍을 통계 이론에 적용하는 방법을 학습한다. 공개된 농산물 관련 빅 데이터를 활용해 농산물 가격을 예측한다. 농산물 가격을 직접 제작한 오픈 플랫폼(웹 사이트)을 통해 게시한다.
연구방법	데이터를 분석하고 수집하는데 유용한 단순 회귀 분석, 다중 회귀 분석, 변수 선택, 가설 검정 같은 통계 이론을 배움과 동시에 능형 회귀 분석, LASSO등 최신 통계 기법을 학습한다. 그리고 실제 데이터 분석에 활용하기 위해 R 프로그램 언어를 웹 사이트 cookbook for r(http://www.cookbook-r.com)을 이용해 배우고 이용한다. 국내의 다양한 농산물 관련 빅 데이터를 다운받고, 다음 년도 농산물 가격과 유의미한 상관관계를 갖도록 지리적 요소와 경제적 요소를 고려하여 농산물 관련 데이터를 분석하고 처리한다. 이후 학습한 다중 회귀분석을 포함한 LASSO, Elastic net등 최신 통계 기법까지 이용하여 다음 년도 배추 가격을 예측해 본다. 또한 개발한 모형 검정 방법을 통해 가장 적합한 모형을 찾아낸다. 최종적으로 다음 년도 가격 예측에 가장 적합한 모형을 오픈 플랫폼 형태로 인터넷상에 웹 사이트로 만들어 게시한다.
연구성과	통계 이론 및 최신 통계 기법을 배우고 적용할 수 있게 되었고 통계 프로그램인 R을 사용할 수 있게 되었다. 이러한 활동들을 통하여 빅 데이터를 다루는 능력이 향상되었다. 다양한 통계 기법을 활용함과 동시에 우리가 고안한 모델 선별 방법을 통해 다음 년도의 배추가격을 예측해 내었다. 마지막으로 그 결과를 토대로 누구나 접근할 수 있는 오픈 플랫폼을 만들어 생산한 정보를 게시하였다. 따라서 농민들이 우리가 예측한 다음 년도의 배추가격을 참고하여 그들의 농사 계획 즉, 미래의 경제 계획에 도움이 되기를 기대한다.
주요어 (Key words)	가격 예측, 통계, 빅 데이터, 생산자중심, 오픈 플랫폼, 선형 회귀, LASSO, Ridge, Elastic net, 변수 선택, 모델 선별, 프로그램 R

1. 개요

□ 연구 동기 및 목적

○ 금융수학에 관심이 있는 학생들을 모아 팀을 결성해 관련 분야 연구를 진행하기로 했다. 금융수학을 같이 공부하면서 최근에 많이 이용되는 빅데이터 처리에 대해 알게 되었고, 이를 실생활에 활용해 보자는 점에서 의견을 같이했다. 그러던 중 우리 학교 근처에서 로컬 푸드 마켓이 열리고 여기서 거래되는 배추 가격이 심하게 변한다는 사실을 알게 되었다. 이를 통계학적 분석 방법을 통하여 가격을 예측하고 농민들에게 알려주면 농산물 재배 계획 수립에 보다 현명한 판단을 돕는 좋은 방안이 될 것이라 생각했다. 또한 이러한 정보를 오픈 플랫폼을 통하여 공개하면 농민들의 의견을 수렴할 수 있고 보다 정확한 가격 예측 정보를 제공할 수 있다. 이를 바탕으로 금융수학과 빅 데이터 처리를 이용한 농민을 위한 오픈 플랫폼을 구축하여 농민들과 소비자가 겪고 있는 배추 가격 변동 문제를 해결하고자 한다.

□ 연구범위

○ 연구 분야

-통계학을 기반으로 농산물의 가격을 예측하였다. 이 과정에서 프로그램 R을 이용하였고 경제학적 관점에서 고민하였다. 이에 본 연구는 수학, 프로그래밍, 경제를 모두 다루고 있다.

○ 연구 범위

- 연구를 진행할 때 필요한 지식과 이론인 통계학을 습득함으로써 배추 가격 예측이 가능하게 되었다.
- 데이터를 통해 모델을 만드는 방법으로 프로그램 R을 익혀 최신 통계 기법을 동원하여 다양한 모델을 만들어 가격을 예측하였다.

○ 연구 진행 단계

- 기존의 농산물 가격 예측 관련 논문들과 통계청 등의 웹사이트를 참고하여 농산물 가격 예측에 대한 선행 연구 조사를 진행하고 농산물 가격을 결정하는 요인들을 선별한다.
- 통계학을 공부하여 실제 분석할 때 필요한 통계적 수학 이론을 습득하고 통계 프로그램인 R을 익혀 다중회귀분석을 포함한 다양한 방법을 통해 모델을 만들고 가격을 예측한다.
- 예측한 가격을 공개하기 위한 오픈 플랫폼(웹 사이트)을 구축한다.

2. 연구 수행 내용

□ 이론적 배경 및 선행 연구

○ 선행 연구

- 박영구 외(2013)는 “배추 무 예측모형 고도화 방안”에서 중요한 작물인 배추와 무의 재배지역을 세분화했고, 작물의 유형별로도 세분화하기 위해 연도별, 시도별 데이터를 사용했다. 또한 고랭지 배추의 수요의 변화경향을 분석함으로써 김치와 수입이 고랭지 배추의 수요에 어떠한 영향을 미치는지도 밝혀내었다.
- 김태훈 외(2014)는 “고랭지 채소면적 변동요인과 전망”에서 실질경영비를 최초로 사용하였으며, 상대수익성 개념을 도입하였지만, 이를 수치화하지 못하였다. 더불어 고랭지 채소가 김치수입 등 수입량에도 영향을 받는다고 조사하였으며, 이를 통해 재배면적의 감소를 추측하였다.
- 심송보 외(2006)는 “농업관측 품목모형 KREI-COMO 2005 개발 운용”에서 전 가격, 공급량, 노임을 재배면적 변인으로, 가격, 공급량, 소득을 가격신축성함수 변인으로 사용하고 국내, 국제가격, 관세, 환율 등을 변인으로 수입수요함수까지 예측해 내었다. 그러나 노임만을 사용한 점, 상대수익성 등 여러 항목들을 고려하지 않고 모든 작물, 지역을 한정했다는 한계가 있다.
- 이용선 외(2012)는 “주요 채소 가격의 변동 패턴 및 요인 분석”에서 추세성, 계절성 등의 시계열적 변동 특성을 활용하였다. 또한 이런 시계열을 활용해 APC, CV등의 지표를 만들고 ARIMA, GARCH등의 모형을 응용하였다. 이를 통해 품목별 변동 특성을 찾아내었고, 변동성이 약 50%정도를 크게 나타나니 이를 고려하여 판단해야 한다고 주장하였다.
- 한석호 외(2010)는 “농업부문 전망모형 KREI-KASMO 2010운용 개발 연구”

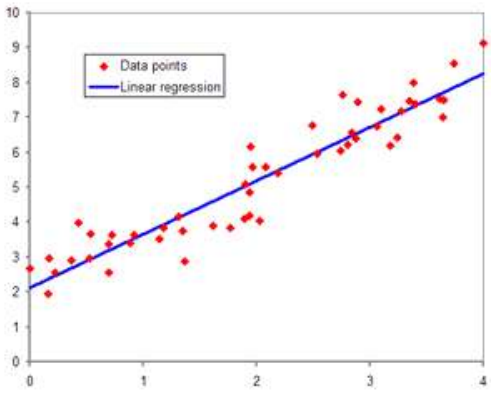
에서 심송보 외(2006)는 “농업관측 품목모형 KREI-COMO 2005 개발 운용”의 모델에서 주요 거시경제 지표 및 FTA에 따른 정책적 영향을 변수에 추가하였다. 따라서 이러한 정책들이 농산물 시장에 미치는 영향을 예측할 수 있도록 하였다.

- 기존의 모형 연구들은 정부 차원에서 가격 및 수급안정을 위한 생산과 유통개선방안 연구가 주를 이루고 있으며, 기후와 지역 차원에서 품목별 변동 특성, 기후변화에 따른 단수의 세분화 등을 다루는 연구도 일부 수행되었다. 하지만 경영비라는 부대비용을 영향이 없다고 판단하거나, 결과가 나와도 정부기관의 웹 사이트에 분기별로 보고서 형식으로 게시하는 등 접근성에 대한 접근은 전혀 시행되고 있지 않다. 또한 농민의 입장을 우선적으로 고려하는 연구는 시행되지 않았었다. 따라서 우리는 농민의 관점으로 경영비와 수익률을 변수에 추가하고, 실질적인 가격 예측에 집중하며, 능형회귀를 포함한 최신 통계기법을 사용해 실질적으로 농민을 위한다는 점에서 차별성이 있다.

	기존 연구	본 연구
경영비	노임만 조사 혹은 영향이 없다고 분석	경영비 외의 수익 등 농민중심 분석
접근성	정부기관의 분기별 보고서 항목	미래 가격 중심의 웹 사이트
변수선정	다단계로 세분화의 각각의 변수를 예측하는 것을 목적으로 함	실질적인 가격을 예측하는 것을 중심으로 함
목적	국가적 계획	농민
모형	통계적 함수 추정, 단순회귀	통계적 함수 포함 능형회귀

○ 단순 회귀 분석(simple regression)

단순 회귀 분석을 통하여 하나의 독립변수와 한 개의 종속변수의 관계를 알아보거나, 종속변수의 값을 예측할 수 있다.



$$y = a + bx + \epsilon$$

위의 결과를 얻을 수 있다. 이것은 종속변수를 예측하는데 결정적인 역할을 하나, 한 번에 한 가지 분석만 가능한 단점이 있다.

◯ 다중 회귀 분석(multiple regression analysis)

독립변수가 두 개 이상일 때 사용하는 회귀 분석이다. 다만 각각의 독립 변수가 모두 독립이여야 한다는 가정이 있다. 각각의 종속 변수와 독립 변수는

$$\begin{aligned}
Y_1 &= a + b_1X_{11} + b_2X_{12} + \cdots + b_nX_{1n} + \epsilon_1 \\
Y_2 &= a + b_1X_{21} + b_2X_{22} + \cdots + b_nX_{2n} + \epsilon_2 \\
\vdots \quad &\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
Y_p &= a + b_1X_{p1} + b_2X_{p2} + \cdots + b_nX_{pn} + \epsilon_p
\end{aligned}$$

이렇게 표현 되고 이는 다시 행렬로

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1n} \\ 1 & X_{21} & X_{p2} & \cdots & X_{pn} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{p1} & X_{p2} & \cdots & X_{pn} \end{pmatrix} \begin{pmatrix} a \\ b_1 \\ \vdots \\ b_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{pmatrix}$$

이렇게 정리될 수 있다. 이때 각각의 b 값은

$$\begin{aligned}
&\sum_{i=0}^{p} (Y_i - a - b_1X_{i1} \cdots b_nX_{in})^2 \\
&= (Y - XB)^T (Y - XB) \\
&= \| Y - XB \|^2
\end{aligned}$$

이렇게 차이의 제곱을 최소화하는 값이다. 이를 구하면,

$$\begin{aligned}
L(B) = L(a, b_1 \cdots b_n) &= \| Y - XB \|^2 \\
&= \| Y - X(X^TX)^{-1}X^TY + X(X^TX)^{-1}X^TY - XB \|^2
\end{aligned}$$

$H = X(X^TX)^{-1}X^T$라 하자.

$$\| Y - XB \|^2 = \| Y - HY \|^2 + \| HY - XB \|^2 + 2(Y - HY)^T(HY - XB)$$

$$(Y - HY)^T(HY - XB) = Y^THY - Y^TXB - Y^TH^THY + Y^TH^TXB$$

그런데 $H = H^T, HX = X, X^T H = X^T, H^2 = H$

$\therefore (Y - HY)^T (HY - XB) = 0$

$\therefore \| Y - XB \|^2 = \| Y - HY \|^2 + \| HY - XB \|^2$

$\therefore X\hat{B} = HY = X(X^T X)^{-1} X^T Y$

$\hat{B} = (X^T X)^{-1} X^T Y$

$\hat{Y} = HY$

◯ 다중 공산성(multiple regression analysis)

다중 회귀 분석에서는 각각의 변수들이 독립임을 가정한다. 하지만 실제로는 각각의 변수들이 완전한 독립일 수는 없다. 따라서 이에 따른 오류가 발생하는데 이것을 다중 공산성(multiple regression analysis)이라고 한다. 이를 해결하는 몇 가지 기법들이 있다.

◯ 상관 분석(multiple regression analysis)

각각의 두 독립변수 사이의 상관성을 0에서 1까지의 숫자 중 하나로 나타내는 것이다. 구하는 식은 아래와 같다.

$$\rho = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

보통 0.45이상을 '상관이 있다.' 0.7이상을 '유의하다.' 라고 판단한다. 다만 한 번에 두 가지 변수만 볼 수 있다는 한계가 있어, 변수의 개수를 효과적으로 제한하지 못한다.

○ 변수 선택 방법(multiple regression analysis)

다중 회귀 분석에서, 변수를 하나씩 줄이거나, 단순 회귀 분석에서 변수를 하나씩 늘려가며 그 중 가장 정확도가 높은 식을 채택하는 방식이다. 변수가 많을 시에는 변수를 줄이는 방법이 가장 적합하다.

-모든 가능한 회귀 (All possible method)

변수가 k개라면 2^k-1개의 모형 중 가장 오차제곱합을 최소로 하는 모형을 찾는 방법이다. k가 큰 경우 현실적으로 사용하기 어렵다.

-전진선택법 (Forward selection)

모형을 구축한 후 모형에 없는 변수들을 하나씩 변수에 넣는 방법이다.

-후진소거법 (Backward elimination)

k개의 변수를 모두 사용해 모형을 구축 한 후 하나씩 줄여가며 오차제곱합을 최소로 만드는 방법이다.

○ 능형 회귀 분석(ridge regression)

Hoerl and Kennard (1970)이 처음 고안한 방법으로 다중 회귀 분석에서, 변수가 데이터의 개수보다 많은 경우를 해결하기 위해서 처음 고안되었다.

$$\hat{B}= (X^TX)^{-1}X^TY$$

<다중회귀분석에서의 최소제곱추정량>

하지만 변수의 개수가 데이터의 개수보다 많은 경우에는

$$(X^TX)\hat{B}= X^TY$$

다음 식을 푸는데 해의 개수가 무수히 많아지게 된다. 따라서

$$\hat{B}= \mathrm{arg}min(\sum_{i=0}^{p}(Y_i-a-b_1X_{i1}-\cdots-b_nX_{in})^2+\lambda_2\sum_{j=1}^{n}b_j^2)$$

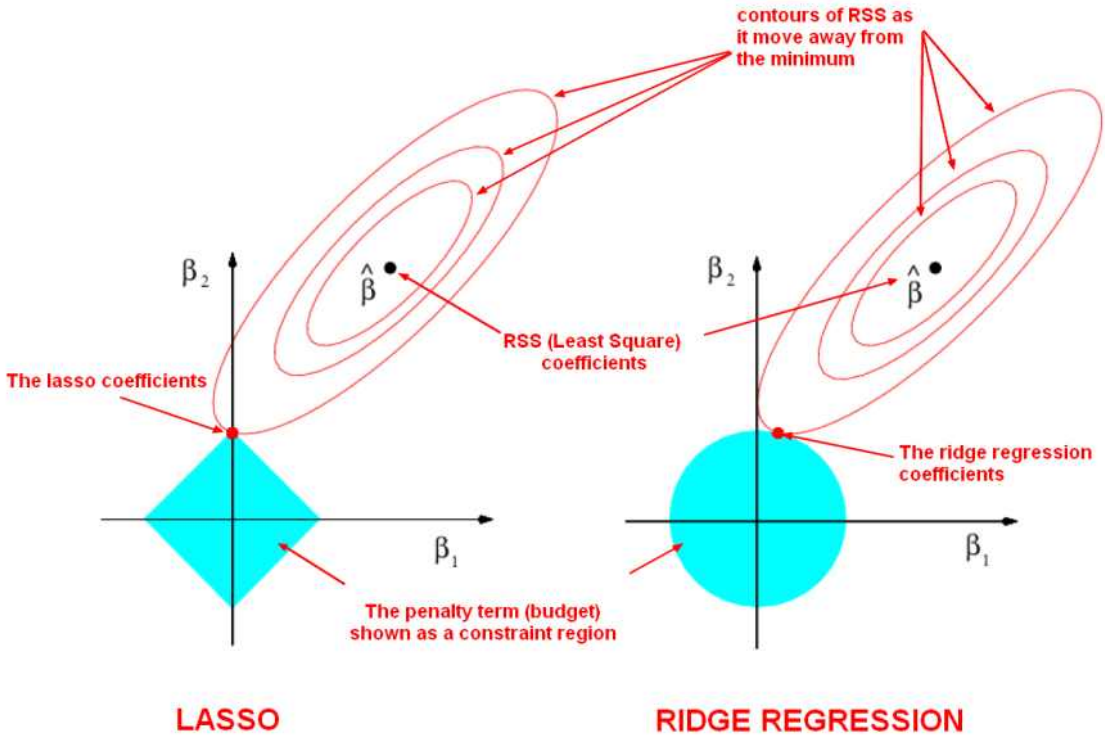
위와 같이 제곱 부분을 추가하고 변형하면 변수의 개수가 더 많은 경우를 해결할 뿐만 아니라 다중공산성을 줄이는 결과가 나올 수 있다. 하지만 모든 변수들이 약간의 보정만을 거친 채로 모형에 포함되어 있어 모형에 대한 해석이 용이하지 않다는 단점이 있다.

○ LASSO regression

Tibshirani (1996)은 위의 능형 회귀 분석에서의 단점을 개선하는 방법을 제안했다. 제곱 부분을 절댓값 부분으로 바꾸면, 다중공산성 감소를 위한 변수 선택과 예측력 향상을 동시에 가져올 수 있다고 한다. 식은 아래와 같다.

$$\hat{B} = \arg min\left(\sum_{i=0}^{p}\left(Y_i - a - b_1 X_{i1} - \cdots - b_n X_{in}\right)^2 + \lambda_1 \sum_{j=1}^{n} |b_j|\right)$$

ridge와 LASSO에서 λ_1, λ_2를 계산하는 방법을 간단히 표현하면 아래와 같다.



출처[Robert Tibshirani(1996)]

○ Elastic net regularization

Trevor (2005)는 2005년에 ridge와 LASSO를 동시에 이용하는 방법을 고안했는데 간단히 표현하면 아래와 같다.

$$\hat{B} = \arg min\left(\sum_{i=0}^{p}\left(Y_i - a - b_1 X_{i1} - \cdots - b_n X_{in}\right)^2 + \lambda_1 \sum_{j=1}^{n} |b_j| + \lambda_2 \sum_{j=1}^{n} b_j^2\right)$$

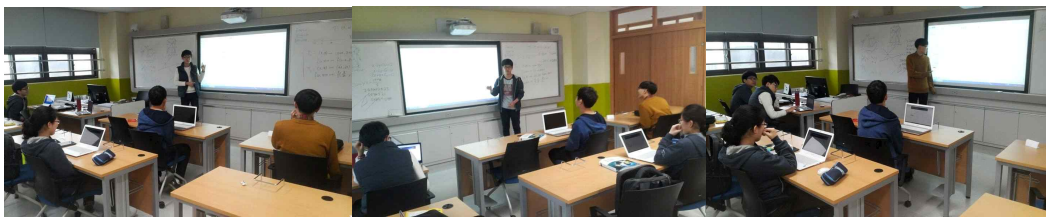
○ 우리는 다중공산성을 줄이는 방법을 5가지 조사하여 모두 사용하였고, 그 중 우리가 고안한 모형평가를 통하여 가장 정확도를 높이는 방법을 선택하였다.

□ 연구주제의 선정

○ 연구주제의 선정과정

- 금융수학과 빅 데이터 처리는 금융 분야에서 불확실한 것들을 예측할 때에 많이 이용되고 있다. 이러한 금융 기술들을 이용하여 실생활에서의 불확실한 부분을 해소한 사례들이 최근 들어 점점 늘어나고 있다. 그래서 우리는 가격을 예측해 보기로 하였고, 예측 대상은 많은 사람들이 심한 가격 변동 때문에 혼란을 겪고 있는 요인들 중 하나인 농산물로 결정하였다. 농산물의 가격은 자연의 영향을 많이 받기 때문에 변동이 심하지만, 그 부분을 금융수학과 빅 데이터 처리를 이용하면 농산물 가격을 예측할 때 오차를 줄일 수 있을 것이다.

우리나라 사람들은 김치와 밀접한 관계를 맺고 있다. 김치의 주재료가 되는 배추 가격에 대해 조사해보니 2010년에는 태풍 곤파스로 인해 배추 가격이 2배로 폭등하였고, 2011년에는 전년도 가격의 급증으로 경작지가 증가하여 다시 배추 가격이 하락하는 등 가격 변동이 심하였다. 심한 배추 가격 변동에 많은 사람들이 큰 피해를 보고 있다는 사실을 알게 되어 통계학과 같이 금융수학에서 이용되고 있는 학문적인 부분들을 공부한 후 금융 기술과 빅 데이터 처리를 익혀 배추 가격을 예측해서 금융 분야만이 아닌 실생활에서 존재하는 불확실성 때문에 사람들이 겪는 피해를 줄이려고 한다.



□ 연구 방법

○ 자료 수집

-재배면적

전년도의 재배면적이 많을수록 가격이 떨어져 농민들이 다음해의 농작물을 결정할 때 다른 작물을 더 고려하기 때문에 변수로 포함하였다.

-국민총소득

국민총소득은 한나라의 국민이 일년 동안 벌어들이는 수익을 계산한 것이다. 국민들의 소득도 구매력에 영향을 미쳐 결국에는 가격의 한 축인 수요에 영향을 미칠 것이라고 생각하여 이를 변수로서 포함하였다.

-수입량

배추의 수입량은 순수한 배추의 수입량과 배추가 약 중량의 50%를 차지하는 김치의 수입량으로 구성되어진다. 따라서 김치수입량과 배추수입량을 모두 고려해 수입량 데이터를 구하는 식을 만들었다. 우리나라의 경우 수입량이 크기 때문에 전체공급량을 고려해주기 위해 변수로서 포함했다.

(수입량) = (배추수입량) + (김치수입량) × (김치중량중배추가차지하는비율)

<수입량을 구하는 식>

-생산량

공급량이 경제학에서 가격을 결정하는데 직접적인 영향을 미치기 때문에 이 변수를 포함했다.

-태풍

배추는 농산물이기 때문에 기후에 많은 영향을 받을 수밖에 없는 특성을 지니고 있어 이 변수를 포함했다. 국가태풍센터에서 제공한 '태풍 발생현황'에서 우리나라에 유효하게 영향을 미친 태풍의 횟수를 활용했다.

-경영비

경영비는 배추농사를 짓는데 있어서 비용의 측면으로 고려된다. 경영비의 상승과 하락은 곧 배추농사 비용의 상승과 하락이다. 그렇기에 경영비의 변화는 향후 배추농사의 공급과 가격에 영향을 미치기에 포함했다.

-기온과 강수량

기온과 강수량은 배추 가격을 결정짓는 가장 중요한 요인이다. 그런데 기상요인은 지리적 위치에 따라 다르기 때문에 지역별 가중치를 고려했다. 다만 배추 가격과 선형적인 관계를 이룬다는 보장이 없어 포함한 모델과 포함하지 않은 모델로 나누어 고려하였다.

$$\frac{(\text{지역}1\text{의생산량}) \times (\text{지역}1\text{의기상요인}) + \cdots + (\text{지역}n\text{의생산량}) \times (\text{지역}n\text{의기상요인})}{(\text{전체지역의생산량})}$$

$= (\text{기온또는강수량데이터})$

<기온, 강수량 데이터 구하는 식>

-출처와 가공여부를 포함한 표

	재배면적	GNI	수입량	생산량	태풍	경영비	기온/ 강수량
출처	통계청	KOSIS 국가통계포털	국세청	통계청	국가 태풍센터	농촌 진흥청	국가기후 데이터센터
가공여부	X	X	O	X	X	X	O

○ 분석 방법

- R 프로그래밍 언어

R 프로그래밍 언어는 통계 계산을 위한 소프트웨어 환경으로, 통계 소프트웨어 개발과 자료 분석에 널리 사용되고 있다. 통계적인 수학 이론을 적용시켜 결과물을 도출시킬 방법으로 프로그램 R을 채택하였다.

- 분석 과정

프로그램 R 사용법을 익히기 위해 프로그램 R의 다양한 기능을 정리해 놓은 인터넷 설명서인 'Cookbook for R'을 참고하였다. 프로그램 R을 사용하여 모델을 검증하는 데에는 아래와 같은 절차를 따랐다.

a. R 프로그래밍 언어 호환을 위해 데이터를 csv 형식으로 가공
b. R에 데이터 입력
c. 모델 검증을 위해 raw data를 test data와 train data로 분리
d. train data로 분석을 진행하여 test data로 검증함.
e. 위의 c~d 절차를 4번 반복하여 오차를 제곱편차로 구함.
f. 모델 평가

c의 과정에서 test data를 선출할 때는 raw data에 기록된 2002년부터 2012

년까지의 데이터 중 반복해가면서 한 년도만을 선택하였고, train data는 기존의 raw data에서 test data의 여집합으로 잡고 분석을 진행하였다. 선택된 연도는 2009년부터 2012년까지 총 4번 검증과정이 진행되었다.
여기서 과정 d의 경우 총 3개의 분석 방법이 R에서 구현되었으며, 다중회귀분석, LASSO, Elastic Net(LASSO와 능형 회귀 분석의 선형결합)을 각각 사용하여 모델들을 테스트하였다.

□ 연구 활동 및 과정

○ R을 이용한 분석

A. 코드 작성 (모델 검증 설계)

```
data <- read.csv("C:/Users/user/Desktop/R&E/csv/8 PP.csv")
train <- data[-8,]
test <- data[8,]

library(glmnet)
library(lars)
x <- as.matrix(train[, -1])
y <- as.matrix(train[, 1])

newx <- as.matrix(test[,-1])
newy <- as.matrix(test[,1])

fit <- lars(x, y, type="lasso")
summary(fit)
best_step <- fit$df[which.min(fit$RSS)]
predictions <- predict(fit, newx, s=best_step, type="fit")$fit
print(predictions)
```

표 1 LASSO를 이용한 코드

```
library(glmnet)

data <- read.csv("C:/Users/user/Desktop/R&E/RR/G18.csv")

train <- data[-8,]
test <- data[8,]

x <- as.matrix(train[, -1])
y <- as.matrix(train[, 1])

newx <- as.matrix(test[, -1])
newy <- as.matrix(test[, 1])

fit <- glmnet(x, y, family="gaussian", alpha=0.5, lambda=0.001)

predictions <- predict(fit, newx, type="link")

print(predictions)
```

표 2 Elastic Net을 이용한 코드

```
data <- read.csv("C:/Users/user/Desktop/R&E/RR/G18.csv")
lm1=lm(medv~crim+rm, train)
summary(lm1)
lm.full=lm(medv~.,data)
summary(lm.full)
score.conti(lm.full,test$medv,test)
lm.back=step(lm.full,direction="backward")
score.conti(lm.back,test$medv,test)
```

표 3 다중회귀분석을 이용한 코드

데이터 분석을 위해 'Cookbook for R'을 참고하여 다중회귀분석, LASSO, Elastic Net를 각각 이용한 코드를 직접 작성하였다.

B. 모델 생성 과정

앞서 말했듯이 각각의 모델에 대해 4번의 과정을 거쳐 정확성을 판별하였다. 모델에 사용된 변수들은 작년도 가격, GNI, 생산량, 재배면적, 배추수입량, 김치수입량, 태풍 횟수, 수입, 경영비, 월별 기온 및 강수량이 있었고, 이 중에 변수들을 적절히 선택하여 모델들을 다수 생성하였다. 모델은

크게 작년도 가격을 포함하는 모델, GNI를 포함하는 모델, 둘 다 포함하지 않는 세 가지 계열로 나뉘었으며, 각각의 계열 내에서는 기온 및 강수량 데이터에 사용할 월을 정하는 것으로 차별성을 두었다.

C. 분석의 시행착오

I. 변수 가공의 필요성

다중회귀분석을 사용하여 가격을 예측하기 위해서는 각 변수의 독립성이 보장되어야 한다. 즉, 변수의 개수가 주어진 데이터 개수보다 많으면 정상적으로 값을 예측할 수 없다. 이는 마치 연립방정식에서 주어진 변수의 개수보다 식의 수가 많을 때 값이 정해지지 않는 경우와 비슷하다고 해석할 수 있다. 따라서 주어진 데이터의 개수, 2002년부터 2012년까지 11개 안으로 변수의 수를 제한하기 위해 아래와 같은 방법을 사용하였다.

- 전국의 재배면적 및 생산량을 가중치를 두어 가공
- 1년의 기온 및 강수량에서 부분적으로 선택

<자문위원께 오류에 대한 자문을 받고 있는 모습>

○ 월별 추진 실적

주요 활동	시 기	비고
기초 정보 수집(전문가 조언)	15.05.01 ~ 15.05.04	자문위원
자료 수집, 정보 검색	15.05.05 ~ 15.05.11	
조건부 확률 등 확률 기초 이론 습득	15.05.12 ~ 15.06.08	자문위원
가설 검정 등 모델 검증 방법 학습	15.06.09 ~ 15.06.22	자문위원
단순 회귀 분석, 다중 회귀 분석, 프로그램 R	15.06.23 ~ 15.07.20	자문위원
로지스틱 분석, 프로그램 R	15.07.21 ~ 15.08.17	자문위원
LASSO 및 능형 회귀 분석, 프로그램 R	15.08.18 ~ 15.09.14	자문위원
전문가 자문, 검토	15.09.23 ~ 15.10.12	자문위원
배운 프로그래밍과 수학을 바탕으로 오픈 플랫폼 구축 및 사이트 개설	15.10.13 ~ 15.11.01	
정리	15.11.01 ~ 15.11.17	

II. 모델 검증 기준 선택

(i) 제곱편차

모델의 정확성을 판별할 수 있는 여러 가지 방법 중에서 제곱편차의 값을 최소화시키는 방법이 가장 일반적이다. 따라서 4개 연도의 가격을 예측한 후 각각의 차이의 제곱의 합이 최소인 모델을 찾는 것을 목표로 설정하였다.

(ii) 경향성

그러나 단순히 제곱편차가 작다는 단일 기준만을 가지고 모델을 객관적으로 평가하기 힘들다는 판단 하에 '경향성'이라는 기준을 추가하게 되었다. 실제 데이터를 보면 2009년의 배추 가격이 매우 높고, 그 다음 2010년에 급격히 떨어졌다가 다시 2012년까지 완만한 상승세를 보인다. 이러한 경향성을 충실히 반영하여 예측한 모델은 객관적으로 훌륭한 모델이라고 평가할 수 있을 것이다.

D. 검증 과정 및 결과

3개의 분석 방법을 이용한 모델 검증 중 다중회귀분석을 이용한 모델들의 검증 결과는 다음과 같다.

	567	678	789	6	7	8	8PP	9PP	89PP	567PP
2012	1212.654	620.2759	2118.038	882.8343	968.1315	3558.426	751.5768	788.5129	732.9121	275.8673
2011	1246	944.517	1266.099	747.8227	867.5255	854.0968	1115.335	650.6786	2314.165	1131.379
2010	595.3712	1154.014	275.7952	882.0093	652.674	740.7424	990.878	756.0064	863.5599	428.681
2009	781.088	766.9513	805.7569	892.3233	922.604	751.5768	−1260.23	1076.875	776.5729	1539.457
제곱편차	374606.7	322392.7	1976743	48776.11	39456.42	7154360	5539886	48140.02	2300097	825911.6
순위	9	11	5	17	19	1	2	18	3	6
	678PP	789PP	5~9PP	5~10PP	8PPP	9PPP	89PPP	567PPP	678PPP	789PPP
2012	753.3019	732.9121	732.9121	699.7628	792.7127	610.5256	822.2834	926.8142	835.6631	857.1078
2011	1378.969	866.8576	1918.947	1144.384	1076.274	639.4362	967.037	935.0679	932.1977	267.6799
2010	1550.057	949.3835	987.3442	1042.45	905.3002	915.4083	955.8124	332.837	1140.019	995.9233
2009	552.4622	371.9847	383.2101	41.18758	421.5915	696.1643	373.9565	563.0525	743.5383	778.9895
제곱편차	1177735	536362.7	1705472	1249508	496558.6	277820.8	529724	453806.5	248850.9	455485.6
순위	9	15	6	8	18	26	16	20	29	19
	5~9PPP	5~10PPP	M1-6	M1-7	M1-8	M2-6	M2-7	M2-8	M3-6	M3-7
2012	774.3796	783.1794	1052.114	990.1919	1510.947	946.958	844.6111	692.4989	1913.971	1162.194
2011	703.438	997.734	575.989	852.886	1090.13	575.992	868.269	459.951	630.693	857.746

	5	8	3	7	5	7	1		7	9
2010	1196.323	1261.613	756.0075	692.679	682.4211	1094.507	1007.298	1030.692	411.1387	103.803
2009	719.0859	365.1483	518.0702	985.3437	689.5807	768.9327	828.3221	920.6267	433.1239	696.9087
제곱편차	330072.4	763322.7	386844.2	20519.49	581976.6	258759.2	113304.6	271933.9	1593801	650587
순위	16	8	13	29	11	19	25	18	6	9
	M3-8	M4-7	M4-8	M5-6	M5-7	M5-8	M6-6	M6-7	M6-8	기존값
2012	−481.389	689.4216	1482.5	1052.114	811.1475	568.5082	1052.114	695.4016	887.0834	902
2011	723.2427	784.6271	1082.598	580.4051	386.825	900.5227	619.6979	813.1839	618.4437	837
2010	1119.538	1014.71	997.2301	756.0075	900.8727	1187.73	756.0075	969.7093	869.049	774
2009	865.9111	749.5546	776.5278	358.3799	1072.342	818.3727	742.1095	742.1095	664.0629	1062
제곱편차	2084553	203495.9	528624.7	583780.1	227115.3	345778.7	172408.1	183882.2	215377.6	0
순위	4	14	8	7	12	10	16	15	13	

표 4 다중회귀분석을 이용한 모델 검증 결과

3. 연구 결과 및 시사점

□ 연구 결과

○ 통계기법의 차이에 따른 모델 예측 결과

중간연구를 진행할 때까지는 수학적 이론 습득 및 여러 가지 이유로 다중회귀분석만으로 제한된 가격 예측을 할 수 밖에 없었으나, 연구를 확장해나가면서 LASSO와 Ridge Regression 등의 최신 통계 기법을 사용하게 되어 모델의 가격 예측이 더 정확해질 것이라고 기대하였다. 주어진 모델을 기반으로 가격을 예측할 때 LASSO, 그리고 LASSO와 Ridge Regression을 선형적으로 결합한 Elastic Net을 사용하여 총 두 번의 가격 예측이 이루어졌다.

○ 모델 평가 기준

모델의 가격을 예측할 때 총 4개의 연도, 2009, 2010, 2011, 2012년의 가격을 예측하도록 하여 제곱편차가 최소가 되도록 하였다. 모델의 정확성을 예측할 때 제곱편차의 크기 이외에도 가격 변화의 경향성이 현실과 맞는지 확인하였다. 예를 들어 2009년의 경우 나머지 세 개의 연도에 비해 가격이 약 30% 정도 더 높은데, 주어진 모델이 이러한 변칙적인 가격 변화에 대응할

수 있는지 확인하였다.

○ 평가 결과

다중회귀분석으로 분석한 결과, 상위 3개의 모델이 제곱편차가 작으면서 가격의 경향성을 잘 따르고 있다. Model3부터 실제 가격과의 차이가 비교적 크며, Model1과 Model2는 굉장히 높은 정확성을 보이고 있다.

	Model1	Model2	Model3	Model4	Model5	Model6	실제가격
2012	990.1919	968.1315	788.5129	882.8343	844.6111	1052.114	902
2011	852.8867	867.5255	650.6786	747.8227	868.2691	619.6979	837
2010	692.679	652.674	756.0064	882.0093	1007.298	756.0075	774
2009	985.3437	922.604	1076.875	892.3233	828.3221	742.1095	1062
제곱편차	20519.49	39456.42	48140.02	48776.11	113304.6	172408.1	0
순위	1	2	3	4	5	6	

표 5 다중회귀분석을 활용한 모델의 예측 가격 및 오차 (제곱편차 기준 / 상위 6개)

연구 결과, 예상과 다르게 LASSO를 활용한 모델들은 가격을 예측하는 데 있어 정확성 차이가 다중회귀분석에 비해 뚜렷하게 나타나지는 않았다. Elastic Net을 적용한 모델의 경우 예상대로 다중회귀분석보다 매우 뛰어난 정확성을 보였다.

LASSO를 활용한 모델의 경우 작년의 가격과 8월, 9월의 기온, 강수량 데이터를 사용한 모델이 가장 가격을 잘 예측하였으나, 이마저도 2009년에 급상승하는 가격을 예측하지 못했다. 대부분의 모델들이 2009년의 가격을 2010년에 비해 낮게 예측하였다.

	Model1	Model2	Model3	Model4	Model5	Model6	실제가격
2012	787.2781	938.4389	946.4811	814.0824	852.8663	808.127	902
2011	894.353	746.0882	833.6643	991.0748	897.0317	882.0049	837
2010	857.7207	814.8024	761.795	764.0224	871.0221	803.162	774
2009	822.2834	771.2008	753.6552	792.0863	741.2842	733.6805	1062
제곱편차	80923.68	95821.76	97215.17	104421.5	118289.8	119481.7	0
순위	1	2	3	4	5	6	

표 6 LASSO를 활용한 모델의 예측 가격 및 오차 (제곱편차 기준 / 상위 6개)

Elastic Net을 사용한 모델의 경우 가장 정확한 예측을 한 2개의 모델은 국민총소득과 7월의 기온, 강수량을 사용한 모델, 그리고 6월의 기온, 강수

량을 사용한 모델이었다. 또한, 상위 3개의 2개의 모델 같은 경우 2009년에 가격이 급등하는 경향성을 잘 반영하고 있어 가격을 예측하는 데 매우 효과적인 모델임을 알 수 있었다.

	Model1	Model2	Model3	Model4	Model5	Model6	실제가격
2012	986.8161	961.8611	819.8474	796.2339	709.6694	745.471	902
2011	850.0577	865.0042	748.8669	865.4317	784.8617	813.5777	837
2010	692.9526	654.8523	881.5465	935.1599	1013.809	968.7433	774
2009	938.9957	922.0037	886.0773	850.5909	919.0413	755.5209	1062
제곱편차	29063.01	38162.72	57031.54	82661.15	117655	156904.3	0
순위	1	2	3	4	5	6	

표 7 Elastic Net을 활용한 모델의 예측 가격 및 오차 (제곱편차 기준 / 상위 6개)

모든 분석방법을 통합하여 상위 모델 6개를 선출하면 다음과 같다.

	Model1	Model2	Model3	Model4	Model5	Model6	실제가격
2012	990.1919	986.8161	961.8611	968.1315	788.5129	882.8343	902
2011	852.8867	850.0577	865.0042	867.5255	650.6786	747.8227	837
2010	692.679	692.9526	654.8523	652.674	756.0064	882.0093	774
2009	985.3437	938.9957	922.0037	922.604	1076.875	892.3233	1062
제곱편차	20519.49	29063.01	38162.72	39456.42	48140.02	48776.11	0
순위	1	2	3	4	5	6	
분석방법	다중회귀	ElasticNet	ElasticNet	다중회귀	다중회귀	다중회귀	

표 8 모델의 예측 가격 및 오차 (제곱편차 기준 / 상위 6개 / 모든 분석 방법 대상)

분석 결과 다중회귀분석을 사용한 모델이 4개로 가장 많았고, Elastic Net이 2개, LASSO를 이용한 모델은 상위 6개 안에 들지 못할 만큼 정확성이 떨어졌다.

□ 시사점

○ 우리는 이 연구 활동을 통해 최신 통계 기법을 포함한 다양한 통계학 기초 이론을 학습하였으며, 이 통계학 이론들을 적용할 수 있는 무료 프로그램인 R을 사용하는 방법을 익히고 다양한 기법들을 응용해 사용해 유의미한 결과를 이끌어 낼 수 있었다.

또한 이러한 통계적 기법들을 실제 농민들에게 도움이 되겠다는 목적을 갖고 실제로 인터넷 사이트를 개설해 공개하면서 이러한 이론들이 실생활에 어떻게 도움이 되어왔는지, 이론들을 배워 어떻게 도움을 주는지 깨닫게 되었다.

뿐만 아니라, 농민을 위한 다음 년도 가격 예측에 집중 하였다는 점, 그리고 아직까지 농작물 가격 예측에는 사용되지 않았었던 lasso, Elastic net regularization등 통계기법을 사용하여 가격을 예측하였다는 점에서 연구가 의미를 가진다.

4. 홍보 및 사후 활용

□ 우리는 웹페이지를 제작하여 여러 농산물들의 가격을 예측한 것을 오픈 소스로 게재하여 많은 사람들이 이를 볼 수 있게 하고 있다. 지금은 배추가격만을 열람할 수 있게 제작하여 놓았지만 후에 다른 작물도 추가하여 웹페이지를 구성할 예정이다. 홈페이지 주소는 http://sasa.iptime.org 이다.

5. 참고문헌

□ 농촌경제연구원

○ 김태훈, 박지연, 박영구 (2014) “고랭지 채소면적 변동요인과 전망”, 정책연구보고, pp.192

○한석호, 이병훈, 박미성, 승준호, 양현석, 신성철 (2011) “기상요인을 고려한 단수예측모형 개발 연구”, 정책연구보고, pp.152

○이용선, 김종진, 노수정 (2012) “주요 채소 가격의 변동 패턴 및 요인

분석”, 정책연구보고, pp.161
◯박지연, 박영구 (2013) “배추, 무 예측모형 고도화 방안”, 기타연구보고 M125
◯이용선, 심송보 (2006) “농업관측 품목모형 KREI-COMO 2005 개발, 운용”, 워킹페이퍼 W27
◯은상규(2012) “정보엔트로피에 가격변화 확률분포를 적용한 농산물가격 불확정성 계측 모델 개발과 적용”, 석사논문, 서울대학교 대학원
◯농촌진흥청(2015) “농수산물 수급조절, 쉽게 이해하기!”

□ 통계학

○ Efroymson-M. A. (1960) “Multiple regression analysis”, Mathematical Methods for Digital Computers, Ralston A. and Wilf,H. S., (eds.), Wiley, New York.
○ Hocking, R. R. (1976) “The Analysis and Selection of Variables in Linear Regression”, Biometrics, 32.
○ E. Hoerla & Robert W. Kennarda, (1970) “Ridge Regression: Biased Estimation for Nonorthogonal Problems” ,Technometrics pages 55-67
○ Robert Tibshirani(1996) “Regression Shrinkage and Selection via the Lasso” Journal of the Royal Statistical Society. Series B (Methodological)
○ Zou, Hui; Hastie, Trevor (2005). “Regularization and Variable Selection via the Elastic Net” . Journal of the Royal Statistical Society, Series B: 301-320.
○ “Cookbook for R ” http://www.cookbook-r.com/
○ “Download R studio” https://www.rstudio.com/