Internet Search Volume Data as a Predictor of Consumers' Daily Food Consumption

Taeyoung Kim[1], Sung-Yong Kim[2], Kyunsik Lee[3]

1. Dept. of Food and Resource Economics (Inst. of Agri. & Life Sci.), Gyeongsang National University, Jinju, Republic of Korea, tykim74@gmail.com
2. Dept. of Food and Resource Economics (Inst. of Agri. & Life Sci.), Gyeongsang National University, Jinju, Republic of Korea, sungyong@gnu.ac.kr
3. Dept. of Food and Resource Economics, Gyeongsang National University, Jinju, Republic of Korea, lilycard@gnu.ac.kr

Internet Search Volume Data as a Predictor of Consumers' Daily Food Consumption

Taeyoung Kim[1] · Sung-Yong Kim[1] · Kyunsik Lee[2]

Department of Food and Resource Economics (Inst. of Agri. & Life Sci.), Gyeongsang National University, Jinju-si, Republic of Korea
Department of Food and Resource Economics, Gyeongsang National University, Jinju-si, Republic of Korea

Introduction

Need for short-term forecast of food consumption

☑ **Food consumption fluctuates by various factors**
- The demand for food is generally stable, while it can change drastically in the short term because of accidents related to food safety, media reports, unpredictable trends, and social atmosphere.
- When a consumer demand at the retail stage is estimated incorrectly, a bullwhip effect occurs in the food supply chain and increases the variation of supply and price in the upper part of the chain.
- Accurate estimation of the short term food demand is an essential task for those engaged in the food industry, including agricultural producers, food manufacturers, food distributors, and policymakers.

☑ **Challenges and Limitations of Household Survey**
- Official data of the consumer demand for food has been obtained through a survey on households.
- However, it is becoming difficult to conduct such surveys due to an increase in the number of single-person households and dual-earner households as well as the demand for safeguarding personal information.
- Moreover, the process of officially deriving statistical survey results requires several months, thus being inappropriate for short-term estimation.

☑ **Internet search data can be used for quick and accurate decision-making in the short term.**
- Big data is receiving attention as a new type of tool that can enhance survey-based data by capitalizing on the development of information technology (the internet in particular).
- Problems with using big data: Big data (e.g. scanned data in retail stores, web scraping data in online retail stores, and social media data) cannot be easily obtained, as they are possessed by private firms and can be purchased and utilized for decision making only at a high price.
- Internet search data are considered an alternative to such big data, which can be used free of charge in real time, making it easy to use for short-term prediction of food demand.

Literature reviews

☑ **The usefulness of Internet search data proved through numerous studies**
- Studies on applying the data obtained from Internet search engines such as Google for the analysis or estimation of the real economy have been carried out for several years in various fields.
- However, only a few studies have examined the effectiveness of Internet search data for predicting the consumer demand for food at the retail stage.
- Particularly, there is a lack of studies that analyze the effectiveness of Internet search data for swiftly predicting consumers' daily food consumption.

Objectives

☑ **This study analyzes the effectiveness of Internet search data to speedily estimate food consumption in a retail scenario.**
- This study examines the effectiveness or availability of Internet search data for predicting the consumer demand for food through two stages.
- First, it verifies whether the volume of Internet searches for a certain item precedes a change in the actual amount spent by consumers on the purchase. Accordingly, the stability analysis of time series data, cross-correlation coefficient analysis of two time series data, and a Granger causality analysis are carried out.
- Second, the study examines an increase in the predictability of a prediction model by applying the volume of Internet searches as an explanatory variable when the practical quantity of purchase for a certain item is estimated.

Methods

Data

☑ **A consumer panel survey conducted by Korea Rural Development Administration (RDA)**
- **Daily food consumption records** such as purchased items, cost of purchase, quantity purchased, price of purchased unit, and purchase channel
  - Recorded in the transaction receipts of consumer panels between January 2017 and December 2017
- **Socio-demographic characteristics of individual households** such as age, household income, number of household members, number of eating out

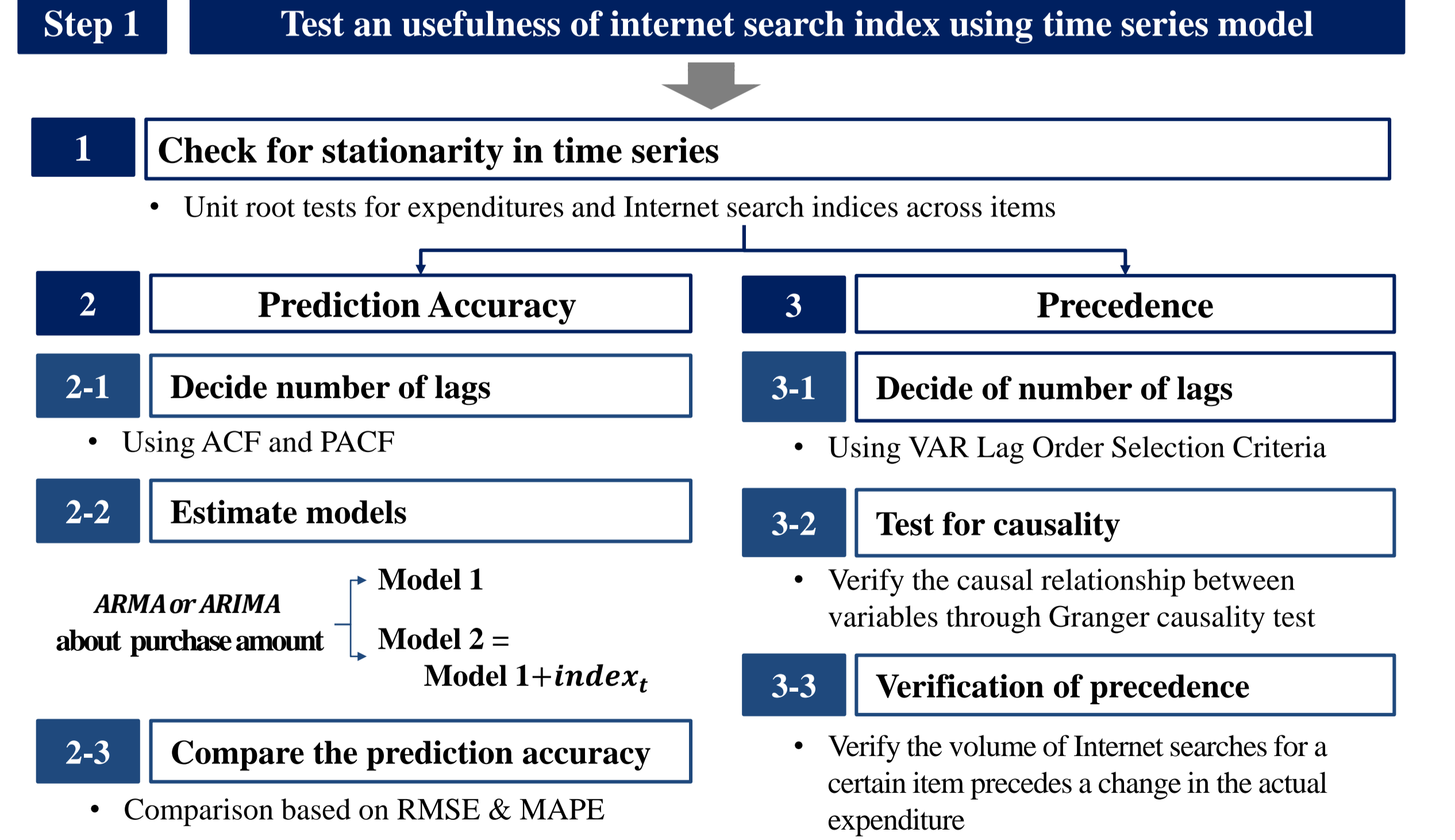☑ **Internet Search Index (ISI) from NAVER search engine**
- **NAVER**, a Korean Internet search website, accounted for 84.03% of the entire volume of Internet searches (85.7% of searches for only food & beverage) in Korea in the past two years (internettrend.co.kr).
- NAVER has been releasing Internet search index (0~100) by date since January 2016.
- Cumulative search index data of the item names and the top 8 most popular search words

☑ **Analysis items**
- Selected items that have high production value or short-term supply and demand fluctuations, and those subject to the government's intensive management of agricultural prices.
- Rice (Korean staple), beef, pork, chicken (Westernized diet, important share of agricultural spending), cabbage, radish, garlic, onion (major spicy vegetables of the season of kimchi)
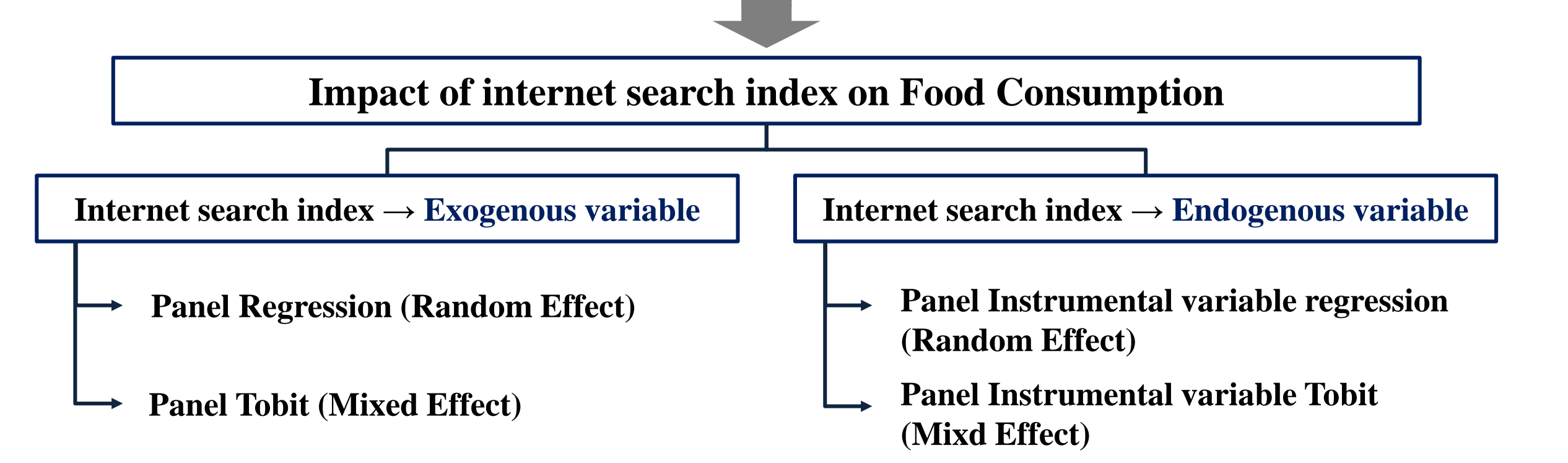
Model selection procedures

☑ **Time series model: used to check prediction accuracy and to verify the precedence of Internet Search Index (ISI).**
- **Prediction accuracy** of the models with/without ISI has been compared based on RMSE and MAPE.
- **Precedence** of ISI over food consumption can verify the usefulness of ISI to elicit the causal relationship between ISI and food consumption

☑ **Panel model: used to elicit the effectiveness of ISI to consumers' food consumption**

Step 1 — Test an usefulness of internet search index using time series model

1. Check for stationarity in time series
- Unit root tests for expenditures and Internet search indices across items

2. Prediction Accuracy

2-1 Decide number of lags
- Using ACF and PACF

2-2 Estimate models

$$\text{ARMA or ARIMA about purchase amount} \begin{cases} \text{Model 1} \\ \text{Model 2} = \text{Model 1} + index_t \end{cases}$$

2-3 Compare the prediction accuracy
- Comparison based on RMSE & MAPE

3. Precedence

3-1 Decide of number of lags
- Using VAR Lag Order Selection Criteria

3-2 Test for causality
- Verify the causal relationship between variables through Granger causality test

3-3 Verification of precedence
- Verify the volume of Internet searches for a certain item precedes a change in the actual expenditure

Step 2 — Test an effectiveness of internet search index using panel model

Impact of internet search index on Food Consumption

Internet search index → Exogenous variable	Internet search index → Endogenous variable
→ Panel Regression (Random Effect)	→ Panel Instrumental variable regression (Random Effect)
→ Panel Tobit (Mixed Effect)	→ Panel Instrumental variable Tobit (Mixd Effect)

Results & Discussion

Time series model: Usefulness of Internet Search Index

1. Check for stationarity in time series

	Expenditures	Internet Search Index
Stationary	Rice***, Beef ***, Pork***, Chicken***, Chinese cabbage*, Radish**, Onion***	Rice***, Beef***, Pork***, Chicken***, Radish**, Garlic***, Onion**
Non-stationary	Garlic	Chicken, Chinese cabbage
1st difference	Garlic***	Chicken***, Chinese cabbage***

***: $p\ value < 0.01$, **: $p\ value < 0.5$, *: $p\ value < 0.1$,
Note: Stationarity test based on Schwarz Criteria

2. Comparison of Prediction Accuracy

2-1 Select number of lags and estimation models

Rice	$ARMA(3,1)$	Chinese cabbage	$ARMA(3,1)$
Beef	$ARMA(1,0)$	Radish	$ARMA(7,0)$
Pork	$ARMA(2,0)$	Garlic	$ARIMA(4,1,1)$
Chicken	$ARMA(3,0)$	Onion	$ARMA(4,2)$

2-2 Comparison of prediction accuracy between the models

	Model 1 RMSE, MAPE	Model 2 RMSE, MAPE		Model 1 RMSE, MAPE	Model 2 RMSE, MAPE
Rice	143.09, 44.30	143.24, 44.33	Chinese cabbage	66.39, 90.25	64.59, 84.12
Beef	1195.83, 55.64	1187.58, 55.05	Radish	45.18, 45.85	44.89, 45.62
Pork	296.31, 29.01	296.45, 29.02	Garlic	75.88, 100.06	75.00, 101.51
Chicken	79.65, 53.09	81.84, 55.52	Onion	28.85, 33.08	28.73, 32.98

3. Check for Precedence

3-1 Decision of number of lags

Rice	6 days	Chinese cabbage	8 days
Beef	1 day	Radish	3 days
Pork	2 days	Garlic	7 days
Chicken	7 days	Onion	2 days

Note: VAR order selection based on Schwarz Criteria

3-2 Test for causality

Internet search index → Purchase amount	Rice, Chicken, Chinese cabbage, Radish, Garlic
Internet search index → Purchase amount / Purchase amount → Internet search index	Onion
No signs	Beef, Pork

3-3 Verification of precedence

Rice	Search index 3 days ago (+) Cost of purchase	Chinese cabbage	Search index 6 days ago (-) Cost of purchase
Beef	–	Radish	Search index 3 days ago (+) Cost of purchase
Pork	–	Garlic	Search index 3 days ago (+) Cost of purchase
Chicken	Search index 5 days ago (+) Cost of purchase	Onion	Cost of purchase 2 days ago (+) Cost of purchase

Panel model: Effectiveness of Internet Search Index

4-1 Check endogeneity of ISI: Test the Validity of Instrumental Variables

☑ ISI can be an endogenous variable with possible IVs such as price & quantity of each item, and seasonal effects.

	IV	Endogeneity test $F-statistic$	Weak identification test $Chi.sq$	Over-identification test $Chi.sq$	Pass or Unpass
Rice	P, Season	1.961	7717.01***	24.27***	X
Beef	P, Q, Season	16.31***	8304.03***	3.83	O
Pork	P, Q, Season	0.907	5863.04***	19.99***	X
Chicken	P, Season	3.14*	11180.08***	7.529*	O
Chinese Cabbage	P, Q, Season	58.99***	765.33***	173.50***	O
Radish	P, Q, Season	9.937**	11676.06***	104.88***	O
Garlic	P, Q, Season	0.509	8042.47***	19.76***	X
Onion	P, Q, Season	4.528**	676.46***	28.56***	O

***: $p\ value < 0.01$, **: $p\ value < 0.5$, *: $p\ value < 0.1$,
P and Q indicate retail price of each item and incoming quantity (or product weight of livestock) at wholesale market, respectively.

4-2 Estimation results

☑ ISI is significantly and positively affect to consumers' food consumption in the most cases.
- ISI can be an effective measure to estimate consumers' food consumption

	Panel Regression (Random Effect)	Panel IV Regression (Random Effect)	Panel Tobit (Mixed Effect)	Panel IV Tobit (Mixed Effect)
Rice	339.68***	-	495.11***	-
Beef	545.57***	503.72***	726.65***	663.12***
Pork	-0.86	-	-22.05	-
Chicken	76.26***	78.74***	178.38***	191.29***
Chinese Cabbage	51.12***	111.84***	94.72***	144.47***
Radish	84.12***	90.08***	87.66***	90.48***
Garlic	299.79***	-	241.32***	-
Onion	0.08	-8.23***	-15.36***	-29.70***

***: $p\ value < 0.01$, **: $p\ value < 0.5$, *: $p\ value < 0.1$
Note: Panel regression models use only non-zero cost of purchase as a dependent variable.

Conclusion

☑ **ISI can improve the prediction accuracy of the model, and can play a role as a useful measure to predict a short-term food demand**

☑ **ISI affects positively to consumers' food consumption regardless of its endogeneity.**

☑ **ISI can be a useful measure to build the marketing strategies of food industry, and can also contribute to stabilization of supply and demand of government.**