Yonsei Department of Statistics and Data Science

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

컨퍼런스 순서

[포닥 발표 세션]

- 박준희: 08:40 ~ 09:00
- Feri Setiawan: 09:00 ~ 09:20

[박사 발표 세션1]

- 김규현: 09:20 ~ 09:50

[초청연사 발표 세션]

- 서울대학교 정성규 교수: 10:00 ~ 10:40
- 고려대학교 신승준 교수: 10:40 ~ 11:20
- 성균관대학교 이경재 교수: 11:20 ~ 12:00

[박사 발표 세션2]

- 김수민: 13:15 ~ 13:30
- 박주영: 13:30 ~ 13:45
- 박진아: 13:45 ~ 14:00
- 전예슬: 14:00 ~ 14:15
- 김예은: 14:15 ~ 14:30
- 김지완: 14:30 ~ 14:45
- 고동영: 14:45 ~ 15:00

[석사 발표 세션]

- 임수린: 15:15 ~ 15:27
- 임지원: 15:27 ~ 15:39
- 조세근: 15:39 ~ 15:51
- 정현묵: 15:51 ~ 16:03
- 김은정: 16:03 ~ 16:15
- 박인서: 16:15 ~ 16:27
- 장백준: 16:27 ~ 16:39
- 박지원: 16:39 ~ 16:51
- 엄혜진: 16:51 ~ 17:03
- 오승미: 17:03 ~ 17:15
- 김민규: 17:15 ~ 17:27
- 정여진: 17:27 ~ 17:39
- 김현주: 17:39 ~ 17:51
- 이해환: 17:51 ~ 18:03

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

박준희

- BK21 포닥 과정

베이지안 접근 방식을 이용한 다지역 임상시험을 위한 계층적 선형 모델의 빈도주의적 1종 오류 통제

Abstract : 다지역 임상시험은 단일 프로토콜에 따라 여러 지역에서 동시에 수행되는 임상시험으로 여러가지 장점이 있는 신약 개발 전략입니다. 이러한 장점의 상당 부분은 다지역 임상시험의 계층적 데이터 구조에서 비롯됩니다. 본 발표에서는 이러한 계층적 구조를 더 잘 반영하기 위하여 개발된 Kim and Kang(2020)의 계층적 선형 모델에 기반하여 베이지안 접근 방식을 사용하여 지역 수가 작은 경우에서의 빈도주의적 1종 오류 통제 방법을 다룹니다. 조금 더 구체적으로, prespecified Bayesian credible level을 설정하여 1종 오류를 통제하는 전략을 제시합니다. 시뮬레이션 결과에서 제안한 방법이 1종 오류를 잘 통제함을 보여주었습니다. 또한 민감도 분석을 통해 지역 간 변동성에 대한 정확한 정보를 제공하였을 때, 그리고 여러 유형의 사전 분포를 사용할 때 각각이 결과에 미치는 영향을 조사하였습니다.

시간: 08:40 ~ 9:00

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

Feri Setiawan

- BK21 포닥 과정

Sequential Inter-hop Graph Convolution Neural Network (SIhGCN) for Skeleton-based Human Action Recognition

Abstract : keleton-based human action recognition has attracted a lot of attention due to its capability and potential to provide more information than just using the sequence of RGB images. The use of Graph Convolutional Neural Network (GCN) becomes more popular since it can model the human skeleton very well. However, the existing GCN architectures ignore the different levels of importance on each hop during the feature aggregation and use the final hop information for further calculation, resulting inconsiderable information loss. Besides, they use the standard Laplacian or adjacency matrix to encode the property of a graph into a set of vectors which has a limitation in terms of graph invariants. In this work, we propose a Sequential Inter-hop Graph Convolution Neural Network (SIhGCN) which can capture salient graph information from every single hop rather than the final hop only and our work utilizes the normalized Laplacian matrix which provides better representation since it relates well to graph invariants. The proposed method is validated on two large datasets, NTU-RBG+D and Kinetics, to demonstrate the superiority of our proposed method.

시간: 9:00 ~ 9:20

Yonsei Department of Statistics and Data Science

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

김규현

- 박사 10학기

Smoothed quantile regression for censored residual life

Abstract : We consider a regression modeling of the quantiles of residual life, remaining lifetime at a specific time. For estimation of regression parameters, we propose an induced smoothed version of the existing non-smooth estimating equations approaches. The proposed estimating equations are smooth in regression parameters, so solutions can be readily obtained via standard numerical algorithms. Moreover, the smoothness in the proposed estimating equations enables one to obtain a robust sandwich-type covariance estimator of regression estimators aided by an efficient resampling method. To handle data subject to right censoring, inverse probabilities of censoring are incorporated as weights. The consistency and asymptotic normality of the proposed estimator are established. Extensive simulation studies are conducted to verify performances of the proposed estimator under various finite samples settings. We apply the proposed method to dental study data evaluating the longevity of dental restorations.

시간: 09:20 ~ 09:50

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

정성규

- 초청연사 강연
- 서울대학교 교수

Double data piling and negatively ridged classifiers in high dimensions

Abstract : Data piling refers to the phenomenon that training data vectors from each class project to a single point for classification. While this interesting phenomenon has been a key to understanding many distinctive properties of high-dimensional discrimination, the theoretical underpinning of data piling is far from properly established. In this work, high-dimensional asymptotics of data piling is investigated under a spiked covariance model, which reveals its close connection to the well-known ridged linear classifier. In particular, by projecting the ridge discriminant vector onto the subspace spanned by the leading principal component directions and the maximal data piling vector, we show that a negatively ridged discriminant vector can asymptotically achieve data piling of independent test data, essentially yielding a perfect classification. The second data piling direction is obtained purely from training data and shown to have a maximal property. Furthermore, asymptotic perfect classification occurs only along the second data piling direction. This interesting phenomenon is shown to also occur in multi-category classification problems, in which the second data piling subspaces are estimated by negatively ridged discriminant subspaces. We demonstrate that negative ridge parameters can be optimal in classification of well-known image and microarray datasets.

시간: 10:00 ~ 10:40

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

신승준

- 초청연사 강연
- 고려대학교 교수

Principal Weighted Least Square Support Vector Machine: An Online Dimension-Reduction Tool for Binary Classification

Abstract : As relevant technologies advance, steamed data that is continuously collected are frequently encountered in various applications, and the need for scalable algorithms becomes urgent. In this article, we propose the principal weighted least square support vector machine(PWLSSVM) as a novel tool for SDR in binary classification, in which most SDR methods suffer since they assumes continuous Y. We further show that the PWLSSVM can be employed for the real-time SDR for the streamed data. Namely, the PWLSSVM estimator can be directly updated from the new data without having old data. We explore the asymptotic properties of the PWLSSVM estimator and demonstrate its promising performance in terms of both estimation accuracy and computational efficiency for both simulated and real data analysis.

시간: 10:40 ~ 11:20

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

이경재

- 초청연사 강연

- 성균관대학교 교수

Scalable and optimal Bayesian inference for sparse covariance matrices via screened beta-mixture prior

Abstract : In this paper, we consider a high-dimensional setting where the number of variables p can grow to infinity as the sample size n gets larger. We assume that most of off-diagonal entries of the covariance matrix are zero. Several Bayesian methods for sparse covariance matrices have been proposed, but their computational speed is too slow, making them almost impossible to apply even to moderately high dimensions (e.g., p ≈ 200$). Motivated by this, we propose a scalable Bayesian method for large sparse covariance matrices. The main strategy of the proposed method is as follows: we first safely reduce the number of effective parameters in a covariance matrix, and then impose shrinkage priors only for selected nonzero off-diagonal entries. To this end, we suggest using the sure screening by keeping only the off-diagonal entries whose absolute sample correlation coefficients are larger than a threshold and furnishing the rests with zeros. It turns out that the proposed prior achieves the minimax or nearly minimax rate for sparse covariance matrices under the Frobenius norm. Therefore, it is not only computationally scalable but also optimal in terms of posterior convergence rate.

시간: 11:20 ~ 12:00

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

김수민

- 박사 7학기

한국 HIV/AIDS 감염인의 잔여수명에 대한 분위수 회귀 분석

Abstract : 잔여 수명 예측은 HIV 감염인들 뿐만 아니라 이들을 치료하는 의사들의 주요 관심사이다. 기존의 생존분석은 연구 시작시점에서의 정보를 기반으로 분석과 예측을 하지만 잔여수명(residual lifetime)을 통한 연구는 임의의 시점까지의 추가적인 정보를 더해 분석 시점에 따른 좀 더 역동적인(dynamic) 분석이 가능하다는 장점이 있다. 또한 생존시간이 주로 오른쪽으로 꼬리가 긴 치우친(skewed) 분포를 가지므로 평균 보다는 중위수(median)가 분포에 대한 더 유용한 요약통계량이고, 이 연구에서는 중위수를 특별한 경우로 포함하는 분위수에 대한 모형화를 고려함으로써 HIV 감염인의 잔여수명을 예측한다. 한편 HIV 감염인에게 있어 CD4 세포 수는 HIV 감염인의 면역 혹은 AIDS로의 진행에 대한 중요한 생체지표(biomarker)이다. 지금까지 CD4 세포 수가 HIV 감염자의 사망 시간에 미치는 영향에 대해 많은 연구가 진행되어 왔고, 한국 HIV/AIDS 코호트 연구에서는 이 CD4 세포수가 동일 대상자에 대해 주기적으로 반복 측정되고 있다. 이 연구는 한국 HIV/AIDS 전향적 코호트 연구 분양 자료 전 회차 자료와 후향적 코호트 자료를 활용하여, CD4 세포 수와 같이 반복 측정되는 경시적인 바이오마커(longitudinal biomarker)의 특성을 함께 고려하여 생존 시간이 편향(biased)되게 추정되지 않기 위해 감염인의 잔여수명의 분위수에 대한 회귀 모형을 적합하는 통계적 추론 방법을 제안한다. 생존시간에 대한 회귀모형으로 준모수분위수(SQRL) 모형을 고려하고, 이를 추정하기 위해 점검함수(check function)에 기반을 둔 불연속 추정함수를 유도평활(induced smoothing)한 추정함수를 제안하여 계산의 효율성을 높인다. 생존 시간에 대한 회귀 모형으로 이 연구를 통해 AIDS와 같이 삶을 위협하는 질병의 환자에 대해 예후 예측 및 치료제 결정 시 좀 더 직관적인 정보를 제공 가능하게 함으로써 한국 HIV/AIDS 치료비용과 시간의 절감 효과를 높인다.

시간: 13:15 ~ 13:30

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

박주영

– 박사 5학기

비모수가우시안혼합척도를 이용한 시간종속 가속고장시간모형

Abstract : 생존자료는 관심 사건이 발생할 때까지의 시간 형태로 구성되어 있으며, 주로 코호트연구를 통해 수집된다. 코호트 연구는 관심 사건이 일어날 때까지의 시간에 영향을 미치는 공변량에 관심이 있으며, 그 연관성과 효과를 추정하고자 한다. 이때, 생존시간에 로그변환한 회귀형태의 기법인 가속고장시간모형(accelated failure time model, AFT)은 대표적인 생존자료모형이다. 이는 생존시간에 공변량의 효과를 직접적으로 모형화한다는 장점을 가지고 있다. 이러한 회귀기법에 사용되는 공변량은 시간에 따라 변하지 않는 고정된 값으로 가정을 하는데, 실제 코호트 자료에서는 추적기간동안 공변량이 변하는 '시간종속(Time-depednent)'형태를 자주 접할 수 있다. 따라서 본 연구에서는 시간종속 공변량에 대한 AFT 모형을 효과적으로 추론할 수 있는 통계방법론을 제안하고자 한다. 이때, 기저시간(Baseline failure time)의 분포는 비모수 연속 가우시안 척도 혼합 밀도함수(nonparametric continuous scale mixture density, NCGSM)를 이용한다. NCGSM는 상당히 유연한 분포로써 이를 이용한 추정의 효율성을 여러가지 시뮬레이션 상황을 통해 살펴보고자 한다.

시간: 13:30 ~ 13:45

Yonsei Department of Statistics and Data Science

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

박진아

– 석박통합 6학기

Bayesian adaptive phase I/II clinical trial design with competing risk model in personalized medicine

Abstract : Bayesian adaptive randomization is the process of assigning patients to treatment. The patient will receive better treatment using intermediate data by adaptive randomization. In the Bayesian paradigm, a posterior prediction distribution calculates a possible predictive outcome conditional on observed data. In this paper, we propose a Bayesian adaptive phase I/II clinical trial design with competing risk endpoints to evaluate the patient's performances at different radiotherapy dose levels. The utility function is used to compare the treatment results and assign better treatment considering the radiation sensitivity of the patient. Because radiotherapy is a "double-edged sword", We consider two events at the same time, tumor progression time and normal tissue complication time. We evaluate the performance of our model through simulation and compare with simple equal adaptive randomization and the model ignoring competing risk event.

시간: 13:45 ~ 14:00

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

전예슬

- 석박통합 6학기

Bayesian Convolutional Networks-based Generalized Linear Model

Abstract : Neural networks provide complex function approximations between inputs and a response variable for a wide variety of applications. Examples include a classification for images and regression for spatially or temporally correlated data. Although neural networks can improve prediction performance compared to traditional statistical models, interpreting the impact of explanatory variables is difficult. Furthermore, uncertainty quantification for predictions and inference for model parameters are not trivial. To address these challenges, we propose a new Bayes approach by embedding convolutional neural networks (CNN) within the generalized linear models (GLM) framework. Using extracted features from CNN as informative covariates in GLM, our method can improve prediction accuracy and provide interpretations of regression coefficients. By fitting ensemble GLMs across multiple Monte Carlo realizations, we can fully account for uncertainties. We apply our methods to simulated and real data examples, including non-Gaussian spatial data, brain tumor image data, and fMRI data. The algorithm can be broadly applicable to image regressions or correlated data analysis by providing accurate Bayesian inference quickly.

시간: 14:00 ~ 14:15

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

김예은

- 석박통합 6학기

OCT MRI Image Enhancement and Layer Detection using Supervised Cycle-GAN with curve similarity loss.

Abstract : OCT MRI images are used to judge the progression of glaucoma. In particular, the thickness between the upper and middle layers in the OCT image is an important factor in determining the disease progression. However, the OCT image taken in the past has disadvantages in that the image quality is poor, the clarity is lowered and the boundary between the two layers is not clear. In this study, a new method considering curve similarity loss as supervised loss in supervised cycle-GAN is proposed. This method has the advantage of specifying the characteristics that the generator should focus on while maintaining the learning structure of the existing cycle-GAN. By considering the similarity of thickness between the generated image and the real OCT image of the same patient as a supervised loss, the generator is forced to focus on creating an appropriate layer boundary. With this method, the two tasks of image transformation and detecting boundary were solved in one model. As a result, the OCT images of low-quality were transformed to improved ones with the boundary between the two layers.

시간: 14:15 ~ 14:30

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

김지완

– 석박통합 6학기

Hierarchical Generalized Linear Models for ordinal response data in Multiregional Clinical Trials

Abstract : Multiregional clinical trials are conducted in more than one region under a single protocol. The patients in the same region have common intrinsic and extrinsic factors. In other words, the individual patients are nested within their own regions. To demonstrate such structure, hierarchical linear models were proposed for the response variables following a normal distribution by Kim and Kang and hierarchical generalized linear models were proposed for the response variables following the exponential family (Bernoulli distribution and the Poisson distribution). Ordered categorical response including toxicity scale, Rankin Scale(stroke), and Glasgow outcome scale (TBI) is a familiar type of patient outcome in clinical trials. In this article, we consider a response variable following multinomial distribution. We suppose the Hierarchical Proportional Odds model satisfying proportional odds assumption. The proportional odds assumption is a restrictive assumption that is often violated in practice. To address violating the assumption, we suppose Hierarchical Generalized logit model. Simulation study was conducted to investigate empirical powers of supposed models.

시간: 14:30 ~ 14:45

고동영
– 석박통합 5학기

Bayesian nonparametric quantile regression with multiple proxy variables

Abstract : Quantile regression has been widely used as an important statistical methodology in a wide range of applications including economics, biology, ecology, and finance. However, it is well known that the quantile regression can be substantially biased when the covariates are measured with error. In this paper, we propose a new flexible method that produces nonparametric quantile estimation in the presence of multiple proxies of true covariate. Multiple proxy variables become more and more available for an unobserved explanatory variable in regression as the various sources of information become available. Under the classical multiple measurement error assumption of additive error and linear relationship, we combine the multiple proxy variables to enable the inference about quantile relationship between response variable and unobserved regressor. An application study shows that our methodology reveal the unobserved regressor and catch the quantile relationship well for various nonlinear data.

시간: 14:45 ~ 15:00

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

임수린

– 석사 5학기

RHSHBoost : Improving classification performance of imbalanced data

Abstract : In the real world, the class distribution of data is imbalanced. But in many cases, we are interested in rare events. For example, financial fraud or cancelling subscription occurs rarely, but these type of cases are exactly what we want to catch. However, classification of imbalanced data is picky due to the over-representation by majority classes, and it makes the classifier hard to pay attention to minority classes. This study will propose an attractive ensemble classification method called RHSHBoost to address the above problem. This classification method uses random undersampling and weighted ROSE sampling under a boosting algorithm. Within a boosting iteration, XGBoost is used to make a classifier.

시간: 15:15 ~ 15:27

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

임지원

- 석사 4학기

Joint Latent Space Model for Analyzing Egos' Network and Item Responses with Alters' Attributes

Abstract : This paper attempts to model the relationship between Mexican immigrants' network and dental conditions using a joint latent space model. Compared to American-born Latinos, Mexican immigrants have less access to medical care and have considerable difficulty accessing essential dental services. Therefore, it is important to understand how the social network of Mexican immigrants affects the use of oral health services. This paper used data from Tala Survey Study and contains data from 332 Mexican immigrants (ego) who migrated from Mexico to the Midwest of the United States and 1,292 people (alter) who discuss important matters with ego. The ego forms a network and is composed of undirected graphs. Each ego filled out a questionnaire with 42 questions related to dental health, including the presence or absence of medical insurance and medical use related to oral diseases. A questionnaire was also conducted on the attributes of the alter that contains nine attributes which not only contains demographic information such as gender and race, but also contains whether or not to talk frequently about dental matters, and the degree of knowledge about dental conditions. This paper intends to use a latent space model that combines the network model of the ego and the item response model, and the item response model on the attributes of the alter. By reflecting the network of ego in the item response model, the joint latent space model can observe the ego, item response of the ego, and the attributes of the alter in one latent space. This paper aims to examine the interaction between the ego and the alter by applying the network of 332 Mexican immigrants (ego), and the attribute data of 1,292 alter to the joint latent space model.

시간: 15:27 ~ 15:39

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

조세근

- 석사 4학기

비모수적 신뢰도 이론을 이용한 군 장교 휴직 예측에 관한 연구

Abstract : 현재 대한민국에서 심각한 저출산 문제로 인해 정부는 출산장려 정책을 지속적으로 개선·보완하여 시행중이며, 일·가정 양립문화의 정착이 완전하게 되도록 노력하고 있다. 군 내부에서도 정책에 따라 군 간부의 육아휴직은 점차 활성화되고 있다. 여기에 국방부는 2021년 9월 2022-2026 국방중기계획을 발표하며 간부 규모는 현재 상비병력의 31.6%인 19만 6000명에서 2026년에는 병력의 40.5%인 20만 2000명까지 늘릴 계획이며, 직업군인을 장기간 활용할 수 있도록 단기의무복무 인원은 감축하고, 중간계급 규모는 지속 확대한다 하였다. 본 연구는 공공데이터포털과 국방통계연보의 군 장교 정원과 년도별 휴직현황 자료를 손해보험 통계 기법의 일종인 비모수적 신뢰도 이론(Non-parametric Credibility Theory)에 적용시켰다. 신뢰도 이론은 자동차 손해보험료 산출에 널리 사용되고 있는 통계적 기법으로써 특정 위험군의 과거 평균 손실과 전체 위험군의 과거 평균 손실에 가중치를 두어 다음 해의 보험료를 예측하는 방법이다. 현재 군에서 장교가 휴직을 하면 다른 인원을 해당 보직에 배치하거나, 부대 내 다른 인력이 그 업무를 같이 하고 있기 때문에, 출산장려 정책이 지속적으로 보편화되면 지금의 방식으로 인력을 운용하는데 어려움이 있을 것으로 보이며, 군의 정책에도 변화가 필요할 것으로 보인다. 따라서 위 연구를 통해 차년도 군 장교 휴직인원을 예측하고 그에 따른 군의 정책방향을 결정할 수 있도록 기여하고자 하였다.

시간: 15:39 ~ 15:51

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

정현묵

- 석사 4학기

보건의료 빅데이터를 활용한, 라이프 스타일과 만성 질병을 포함한 국민 건강 및 의료 비용의 관계 분석

Abstract : 건강에 대한 관심이 증가함에 따라 우리나라 국민의 라이프 스타일은 전반적으로 변화했다. 또한, 급격한 고령화에 따라 만성 질병에 관한 예방 및 관리는 개인적, 국가적 관심사가 되었다. 하지만 라이프 스타일과 국민 건강의 상관관계에 대한 국내 선행 연구들의 경우, 표본 수가 적으며 의료 데이터를 포괄하지 못한다는 한계를 내포하고 있다. 따라서 본 연구에서는 민간에 개방된 건강보험공단의 표본코호트DB를 활용하여 표본 수의 한계를 보완하고자 한다. 더불어 의료 진료 내역 데이터를 활용하여 라이프 스타일과 국민 건강의 관계성에 대한 전국민적인 통계적 모형을 도출하고, 라이프 스타일의 변화가 야기할 수 있는 국민조건비용의 증감 모형을 구축하고자 한다.

시간: 15:51 ~ 16:03

김은정

- 석사 4학기

보건의료 빅데이터를 활용한 3대 중대질병의 치료 여정 분석 및 사보험 보장내용과의 비교

Abstract : 국민 보건과 복지에 대한 관심 증가와 의료기술의 발달로 건강 검진이 일반화되면서 동시에 사적 보험시장에서 판매되는 다양한 건강보험상품들의 보장 범위 및 내재된 위험의 측정 역시 매우 중요한 이슈로 떠오르고 있다. 그러나 현재 사보험에서 제공하는 중대 질병(암, 심혈관질환, 뇌질환)에 대한 보장 내용이 실제 진단 이후 사망에 이르기까지의 여정 동안 환자가 경험하는 것과는 괴리가 있으며, 기존 선행연구들은 주로 생존율이나 사망률의 추정 등의 통계적인 추정에 머물고 있어 실제 진단 후 유병 형태 및 치료와 완치, 관리, 재발, 악화/사망 또는 합병증 유발 등 전반적인 중대 질병의 치료 여정에 대한 구체적 이해는 부족한 실정이다. 이에 본 연구에서는 국민건강보험공단의 표본 DB를 사용하여 생존 기간과 생존 경로, 치료여정과 그에 따르는 제반 비용들과 현재 사보험사들의 보험 급부와의 일치여부를 통계적 기법을 통해 확인하였다.

시간: 16:03 ~ 16:15

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

박인서

- 석사 4학기

A Measurement Error Approach for Satellite-based Air Quality Modeling: Using GEMS Level-2 NO2 Products

Abstract : Satellite remote sensing has been widely used to conduct a broad range, real-time monitoring of air quality. Satellite data, however, have potential sources of error in measurement. For example, cloud cover may increase uncertainties or even disable remote sensing, because satellite sensors can be affected by reduced visibility and radiographic distortion. Thus, in the fields of atmospheric science, it is a common practice to discard unreliable pixels by introducing certain thresholds to those sources of measurement error. Although the threshold approach is one way to ensure the quality of satellite data, from a statistical viewpoint, this can cause bias and efficiency loss. Therefore, we propose an alternative method to model satellite-based air quality estimates using a measurement error approach. Our research problem is characterized by the following three properties. First, we aim to estimate the mean fields from satellite observations. Second, it is assumed those observations are prone to measurement error. Third, both observations and errors have spatio-temporal dependence structure. Against this backdrop, we propose a general measurement error framework to model spatio-temporal mean fields with error in the response. For model evaluation, we plan a Monte Carlo simulation study using satellite retrievals of nitrogen dioxide (NO2) from the Geostationary Environment Monitoring Spectrometer (GEMS). The proposed method and existing approaches will be applied to the synthesized NO2 data, and compared on their performances.

시간: 16:15 ~ 16:27

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

장백준

- 석사 4학기

Multiple imputation for continuous regression model with outcome dependent sampling

Abstract : In observational studies, resources may not be available to measure all the information for entire population. In cases where the outcome is relatively easy to measure compared with the exposures of interest, then outcome dependent sampling could be a good alternative. Outcome dependent sampling is a sampling design in which individuals are chosen with probabilities that depend on the observed outcomes. We can save resources and focus it on the parts of the population that are believed to be most informative. If the outcome dependent sampling is used when the outcome variable is continuous, however, standard estimation methods such as maximum likelihood or ordinary least squares assuming iid data, will no longer make consistent, or even unbiased, estimators. In this article, we will describe several estimation methods that solve this problem. And we will show another approach that imputes missing values from outcome dependent sampling. Finally, we can compare the results from the estimation method and the one from multiple imputation method.

시간: 16:27 ~ 16:39

박지원

– 석사 4학기

EM by weighting for capture-recapture data

Abstract : Capture-recapture data which typically includes several times of re-observations remains incomplete due to partially missing covariates. In this paper, EM by weighting is used to impute missing values by assigning weights for an augmented data set with a non-parametric assumption under MAR situation. After calculating a weight for each possible case, we can generate a data set using PPS(Probability proportional to size) sampling. Then, model parameters can be estimated from the complete data set and compared with those of fully observed data. The comparison results are illustrated with simulation studies.

시간: 16:39 ~ 16:51

Yonsei Department of Statistics and Data Science

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

엄혜진

– 석사 4학기

Bayesian Hierarchical Growth-curve modeling of the increasing number of restaurant review during COVID19

Abstract : Time series models are mainly used for the estimation of the data growing over time in hospitality domain. Growth-shape data, however, is not the main interest of the model. On the other hand, Latent growth model is specified for the increasing data estimation, but limited to a linear model. In order to improve modeling for the increasing data by time, growth curve modeling in Bayesian framework is suggested in this paper. Since Bayesian method could be easily applied in hierarchical structure, Bayesian hierarchical model and growth curve model as level1 modeling is used. With the hierarchical model, the effect of independent variables on total growth curve and individual growth curves could be estimated from level2 and level1. For the estimation, flexible Schnute's growth curve is first fitted to the data and specified growth curve model, that is Gompertz in the paper is then applied for the final estimation. The number of restaurants' weekly review in Korea from 2019 to 2021.Feb is estimated in the paper, and the result shows that only delivery affected the growing number of reviews during increasing period. In conclusion, delivery is the key factor of the growth of the restaurants in Korea during COVID19.

시간: 16:51 ~ 17:03

Yonsei Department of Statistics and Data Science

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

오승미

– 석사 4학기

Bayesian Neural Network approach on Infectious Disease Modeling: COVID-19 cases in South Korea

Abstract : After COVID-19 outbreaks, many deep learning based studies are proposed to predict the number of infected individuals. However, deep learning approach focuses solely on improving accuracy with highly over-confidence, which is not appropriate for unprecedented infectious disease cases where reliability in prediction is particularly crucial. In this paper, we examine Bayesian neural networks using Monte Carlo dropouts to forecast the COVID-19 cases in South Korea and estimate uncertainty belonging to a prediction period. We pointwisely estimate transmission rate, recovery rate and reproduction rate based on Susceptible-Infected-Recovered (SIR) model to point out infectiousness of COVID-19. Also, by concatenating predicted confirmed cases and external features and proceeding a dense layer at the last step, we interpret the effects of government actions such as workplace closing, social distancing to reduce the spread based on learned weights. We believe that this study can be applied to other infectious disease cases with higher reliability and interpretability.

시간: 17:03 ~ 17:15

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

김민규

– 석사 4학기

Point Process Cluster Analysis in Latent Space of Item Response Model for School Survey Data

Abstract : We can postulate a process for finding clusters of item latent positions for survey data. The process is motivated by latent space item response model and point process. Respondents and items are embedded in latent space from the model. We can find networks of respondents and items by analyzing the latent space. For analyzing clusters in item latent positions statistically, we apply point process to item response result. Here we use the school survey data of Gyeonggi Provincial Education Office. By fitting point process on item response result, we can detect item clusters of survey data and find differences between school levels and innovative status.

시간: 17:15 ~ 17:27

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스

정여진

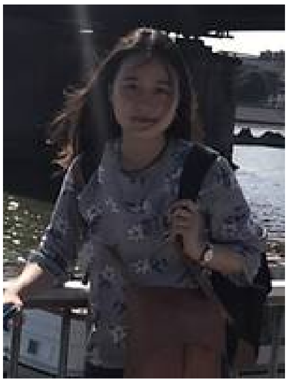
– 석사 4학기

Fast Bayesian Functional Regression for Non-Gaussian Spatial Data

Abstract : Functional generalized linear models (FGLM) have been widely used to study the relations between non-Gaussian response and functional covariates. However, most existing works assume independence among observations and there- fore they are of limited applicability for correlated data. A particularly impor- tant example is spatial functional data, where we observe functions over spatial domains, such as the age population curve or temperature curve at each areal unit. In this paper, we extend FGLM by incorporating spatial random effects. However, such models have computational and inferential challenges. The high- dimensional spatial random effects cause the slow mixing of Markov chain Monte Carlo (MCMC) algorithms. Furthermore, spatial confounding can lead to bias in parameter estimates and inflate their variances. To address these issues, we propose an efficient Bayesian method using a sparse reparameterization of high-dimensional random effects. Furthermore, we study an often-overlooked challenge in functional spatial regression: practical issues in obtaining credible bands of functional parameters and assessing whether they provide nominal coverage. We apply our methods to simulated and real data examples, including malaria inci- dence data and US COVID-19 data. The proposed method is fast while providing accurate functional estimates.

시간: 17:27 ~ 17:39

연세대학교 통계데이터사이언스학과 BK학술컨퍼런스



김현주

– 석사 4학기

Latent Space Accumulator Model for Analyzing Bipartite Networks with Connection Times and Its Applications to Item Response Data

Abstract : Information about the response time has been highlyaccessible from recent educational and psychological testing environment, and variousmethodologies incorporating its information have been developed. However, thereare not many statistical models that analyze the effect of response time onnetwork analysis approach. The response time is expressed as 'connection time'in the network framework, and I propose a novel model, a latent spaceaccumulator model for analyzing bipartite networks that have their connectiontime for each connection-type to estimate the effect of connection time onnetwork structure. To model connection times for each connection-type that ismutually exclusive, the competing risk modeling framework was adopted. Toidentify the effect of time on network structure, I embedded latent spaces intothe competing risk models. The model has been successfully applied to the itemresponse data with response time, which can be regarded as one example ofbipartite networks.

시간: 17:39 ~ 17:51

연세대학교 통계데이터사이언스학과
BK학술컨퍼런스

이해환
– 석사 3학기

Raking and Relabeling for Imbalanced Data

Abstract : We consider the binary classification of imbalanced data. A dataset is imbalanced if the proportion of classes are heavily skewed. Imbalanced data classification is often challengeable, especially for high-dimensional data, because unequal classes deteriorate classifier performance. Undersampling the majority class or oversampling the minority class are popular methods to construct balanced samples, facilitating classification performance improvement. However, many existing sampling methods cannot be easily extended to high-dimensional data and mixed data, including categorical variables, because they often require approximating the attribute distributions, which becomes another critical issue. In this paper, we propose a new sampling strategy employing raking and relabeling procedures, such that the attribute values of the majority class are imputed for the values of the minority class in the construction of balanced samples. The proposed algorithms produce comparable performance as existing popular methods but are more flexible regarding the data shape and attribute size. The sampling algorithm is attractive in practice, considering that it does not require density estimation for synthetic data generation in oversampling and is not bothered by mixed-type variables. In addition, the proposed sampling strategy is robust to classifiers in the sense that classification performance is not sensitive to choosing the classifiers.

시간: 17:51 ~ 18:03