NOWY PROGRAM STUDIÓW
2014/2015

STANDARDOWY SYLABUS PRZEDMIOTU KIERUNKOWEGO/SPECJALNOŚCIOWEGO

Koordynator przedmiotu:	**Wykładowcy uczestniczący w opracowaniu sylabusa:**
dr hab. prof. SGH Ewa Frątczak	1. dr hab. prof. SGH Ewa Frątczak 2. dr Wioletta Grzenda 3. dr Aneta Ptak-Chmielewska 4. dr Iga Sikorska 5. mgr Adam Korczyński

Sygnatura:

Tytuł oferty	ZAWANSOWANE METODY ANALIZ STATYSTYCZNYCH
Ang.	ADVANCED STATISTICAL ANALYSIS METHODS

Część A

Syntetyczna charakterystyka przedmiotu *(około 400 znaków)*:
(opis w jęz. polskim)

Przedmiot obejmuje wybrane zaawansowane metody analiz statystycznych, w tym: uogólnione modele liniowe, modele mieszane, modele i analizy wielopoziomowe, analizę zmiennych jakościowych i modele zmiennych ukrytych, wykorzystanie metod Monte Carlo opartych na łańcuchach Markowa w statystyce bayesowskiej oraz techniki doskonalenia jakości danych, braki danych .
Uogólnione modele liniowe obejmują filozofię estymacji oraz zastosowania do wybranych danych empirycznych.
….
Pierwszy segment zajęć z zakresu analizy zmiennych jakościowych i modeli zmiennych ukrytych prezentuje zagadnienia analizy współzależności zmiennych jakościowych oraz podstawowe metody modelowania zmiennych jakościowych. Drugi segment zajęć przedstawia szczegółowo metodę analizy zmiennych ukrytych w ujęciu poprzecznym i wzdłużnym, jako przykłady modelowania cech o charakterze jakościowym.
Zagadnienia dotyczące metod Monte Carlo opartych na łańcuchach Markowa w statystyce bayesowskiej wykorzystane są do estymacji różnych modeli statystycznych.
Techniki doskonalenia jakości danych prowadzą do poprawy jakości informacji otrzymywanych na podstawie danych, a także decyzji bazujących na pozyskanej z danych wiedzy. Zagadnienia braków danych (missing data) obejmują rodzaje brakujących danych oraz klasyczne i nowoczesne metody imputacji.
Zdobyte umiejętności dają podstawy do realizacji projektów dotyczących analizy danych ubezpieczeniowych, finansowych, telekomunikacyjnych i innych. Podczas praktycznych ćwiczeń stosowane są różne pakiety komputerowe.

(opis w jęz. angielskim)

The course includes selected advanced statistical methods including: generalized linear models, mixed models, multilevel models and analysis, the categorical data analysis, Markov Chain Monte Carlo method in Bayesian statistics and techniques of quality data improvement.
Generalized linear models include philosophy of estimation and application based on empirical data.

….
The first part of the categorical data analysis presents the correlation analysis approach applied to categorical variables and the basic methods of categorical data modeling. Second part focuses on the latent class analysis (LCA) and latent transition analysis (LTA), as examples of modeling categorical variables in order to poses latent information.
The Markov Chain Monte Carlo method in Bayesian statistics is used to estimate different statistical models.
Applying techniques of data quality improvements leads to better quality information obtained from data, as well as better decisions based on knowledge discovered in data. Problem of missing data include: types of missing data and methods of imputation.
The course provides the background to realize projects including data analysis in insurance, financial, telecommunication and other sectors. Computer software systems are used for practical exercises.

Część B

Cele zajęć z przedmiotu:
(opis w jęz. polskim)

Celem zajęć jest przekazanie studentom wiedzy dotyczącej zawansowanych metod analiz statystycznych z wykorzystaniem różnych pakietów komputerowych, w tym systemu SAS, oraz wykształcenie umiejętności praktycznego stosowania tych metod.

(opis w jęz. angielskim)

The objective of the course is to provide students with the knowledge on advanced statistical analysis methods using software systems including SAS and ability to apply these methods in practice.

Efekty kształcenia:
To stwierdzenia określające, co student powinien wiedzieć, rozumieć i/lub potrafić zrobić po zakończeniu okresu kształcenia (w ramach przedmiotu). W tych stwierdzeniach należy używać czasowników w stronie czynnej, odnoszącej się do wiedzy, rozumienia, praktycznego zastosowania, analizy, syntezy, oceny, itp.)

Wiedza	*(opis w jęz. polskim)* **Student powinien być w stanie:** 1. znać filozofię estymacji uogólnionych modeli liniowych 2. znać obszary zastosowania uogólnionych modeli liniowych 3. rozróżniać: rodzaje zmiennych jakościowych i znać miary współzależności 4. umieć kodować zmienne oraz wyróżniać modele służące do ich analizy 5. rozumieć koncepcję analizy tablic kontyngencji oraz estymacji Metodą Największej Wiarygodności 6. rozumieć koncepcję zmiennej ukrytej oraz istotę analizy zmiennych ukrytych w ujęciu porzecznym i wzdłużnym 7. znać i rozumieć podstawowe pojęcia oraz metody statystyki bayesowskiej 8. znać i rozumieć podstawowe pojęcia metod Monte Carlo opartych na łańcuchach Markowa 9. rozróżniać poszczególne kategorie jakości danych oraz posługiwać się miarami jakości danych, 10. umieć stosować techniki doskonalenia jakości na rzeczywistych danych. 11. Znać i rozumieć typy braków danych i ich źródła oraz znać tradycyjne i nowoczesne metody imputacji.

	(opis w jęz. angielskim) **The student should:** 1. know philosophy of generalized linear models estimation 2. know the areas of generalized linear models application 3. … 4. distinguish types of categorical variables and know basic measures of correlation applied to categorical data 5. code categorical variables and know methods of categorical data analysis 6. understand the concept of contingency tables and Maximum Likelihood Estimation 7. understand the concept of latent variable, latent class and latent transition analysis 8. know and understand basic notions and methods of Bayesian statistics 9. know and understand basic notions of Markov Chain Monte Carlo methods 10. distinguish types of categories of data quality and know how to use data quality measures. 11. know how to apply data quality improvements tools on real world databases. 12. Know types of missing data and basic methods of imputation
Umiejętności	*(opis w jęz. polskim)* **Student powinien umieć:** 1. zastosować uogólnione modele liniowe na wybranych danych empirycznych 2. zinterpretować wyniki estymacji uogólnionych modeli liniowych 3. umieć kodować zmienne jakościowe, interpretować podstawowe miary współzależności pomiędzy zmiennymi jakościowym 4. znać podstawowe techniki analizy danych jakościowych 5. przygotować zbiór danych do analizy modeli zmiennych ukrytych w ujęciu poprzecznym i wzdłużnym 6. estymować i weryfikować modele zmiennych ukrytych 7. interpretować wyniki analizy zmiennych ukrytych 8. włączać wiedzę a priori do modelu statystycznego 9. estymować bayesowskie modele statystyczne 10. umieć zidentyfikować typ problemu i zastosować właściwe narzędzia poprawy jakości danych, 11. umieć oceniać jakość danych pochodzących z badań ankietowych oraz określać wpływ błędów na wyniki analizy danych. 12. umieć zidentyfikować rodzaje braków danych i zastosować właściwe metody imputacji
	(opis w jęz. angielskim) **The student should be able :** 1. apply generalized linear models on empirical data 2. interpret results of generalized linear models estimation 3. know how to code and interpret categorical variables as well as interpret basic measures of correlation analysis applied to the categorical data 4. know basic method of categorical data analysis 5. be able to prepare data set for the latent class and latent transition analysis

	6. be able to estimate and verify latent class and latent transition models 7. interpret the output of the latent class models 8. to include prior knowledge in a statistical model 9. to estimate Bayesian statistical models 10. identify the type of problem and use proper tools for data quality improvement, 11. know how to evaluate quality of survey data and assess error effects on results of the data analysis. 12. identify the type of problem of missing data and apply proper methods of imputation.
Inne kompetencje	*(opis w jęz. polskim)* 1. (…)
	(opis w jęz. angielskim) 1. (…)

Część C

Semestralny plan zajęć:
(opis w jęz. polskim)

1. UML idea zastosowania. Założenia oraz główne składowe modelu. Rodzina rozkładów wykładniczych. Funkcja łącznikowa.
2. Podstawowe metody estymacji modelu. Jakość modelu, zasady wyboru modelu optymalnego. Przykłady zastosowań – przykłady estymacji i weryfikacji UML.
3. Zmienne jakościowe, rodzaje, sposoby kodowanie. Podstawowe miary współzależności pomiędzy zmiennymi jakościowymi. Analiza tablic kontyngencji. Estymacja metodą największej wiarygodności.
4. Koncepcja regresji logistycznej jako przykład metody analizy zmiennych jakościowych. Istota analizy zmiennych ukrytych w ujęciu poprzecznym oraz wzdłużnym. Interpretacja parametrów w modelach zmiennych ukrytych.
5. Estymacja i weryfikacja modeli zmiennych ukrytych w ujęciu wzdłużnym i porzecznym. Dodatkowe zagadnienia (zmienne grupujące, zmienne kontrolujące, aplikacja modelu zmiennych ukrytych do modelowania regresji logistycznej o postaci wielomianu).
6. Algorytm estymacji modeli zmiennych ukrytych w ujęciu poprzecznym. Estymacja modelu. Interpretacja uzyskanych wyników.
7. Algorytm estymacji modeli zmiennych ukrytych w ujęciu wzdłużnym. Estymacja modelu. Interpretacja uzyskanych wyników.
8. Idea metod bayesowskich. Podejście klasyczne a podejście bayesowskie. Rozkłady a priori. Rozkłady a posteriori.
9. Wnioskowanie bayesowskie. Estymacja punktowa. Bayesowskie przedziały ufności. Weryfikacja hipotez. Hierarchiczne i empiryczne modele bayesowskie. Porównywanie modeli.
10. Metody Monte Carlo oparte na łańcuchach Markowa. Wybrane własności łańcuchów Markowa. Rozkłady stacjonarne. Twierdzenia ergodyczne. Metody Monte Carlo w statystyce bayesowskiej. Metody Monte Carlo z funkcją ważności.
11. Nowoczesne metody symulacji. Algorytm Metropolisa. Algorytm Metropolisa-Hastingsa. Próbnik Gibbsa. Testy zbieżności łańcuchów Markowa.
12. Przykłady zastosowań metod MCMC w statystyce bayesowskiej w systemie SAS. Bayesowska estymacja modeli regresji liniowej. Bayesowska estymacja wielowymiarowych modeli regresji. Bayesowska estymacja uogólnionych modeli liniowych. Inne modele bayesowskie.
13. Etapy budowy modeli – podejście bayesowskie a tradycyjne. Wybór rozkładów a priori. Zagadnienia dotyczące

wyboru realizacji łańcucha Markowa. Interpretacja wyników.
14. Kategorie jakości danych: jakość wewnętrzna (dokładność, obiektywność, wiarygodność), kontekst (relewantność, aktualność, kompletność), dostępność, reprezentacja (łatwość zrozumienia, zwięzłość, spójność).
15. Kompleksowe zarządzanie jakością danych (TDQM) – zasady, miary jakości danych i narzędzia doskonalenia jakości danych.
16. Wyznaczanie wartości miar jakości danych ze względu na różne kategorie na rzeczywistych zbiorach danych, doskonalenie jakości danych pochodzących z badań ankietowych.
17. Normalizacja i standaryzacja danych, przekształcanie i modyfikacja zmiennych.
18. Stosowanie narzędzi doskonalenia jakości danych, poprawa spójności i dokładności danych, parsing danych.
19. Podstawowe pojęcia w problematyce braków danych. Monotoniczny i ogólny wzorzec brakujących danych. Mechanizmy brakujących danych (braki danych całkowicie losowe, losowe i nielosowe).
20. Wprowadzenie do metod imputacji danych. Metody tradycyjne: usuwanie obserwacji oraz metody imputacji pojedynczej.
21. Współczesne metody imputacji danych. Metody bazujące na estymatorach największej wiarygodności. Imputacja wielokrotna.

(opis w jęz. Angielskim)

1. Generalized linear models – philosophy of application. Assumptions and basic components of the model.Exponential functions family. Link functions.
2. Basic estimation methods. Model's quality and rules of selection of optimal model. Examples of application – examples of estimation and GLM verification.
3. Categorical variables, types and different ways of coding. Correlation measures applied to categorical data. Contingency tables. Maximum Likelihood Method of estimation.
4. Binary logistic regression as an example of categorical data analysis. Latent class modeling concept (LCA and LTA approach). Interpretation of the latent class models results.
5. Estimation and verification of the latent class models (LCA and LTA). Additional aspects (grouping and controlling variables in LCA).
6. Estimation procedure of the latent class analysis. Interpretation of the latent class analysis output.
7. Estimation procedure of the latent transition analysis. Interpretation of the latent transition analysis output.
8. Bayesian method. Bayesian approach vs. classical approach. Prior distributions. Posteriori distributions.
9. Bayesian inference. Point estimation. Bayesian interval estimation. Hypothesis testing. Hierarchical and empirical Bayesian models. Comparison of models.
10. The Markov Chain Monte Carlo method. Some properties of Markov chains. Stationary distributions. Ergodic theorems. Monte Carlo method in Bayesian statistics. Monte Carlo method with importance function.
11. Modern model estimation. Metropolis algorithm and Metropolis-Hastings algorithm. Gibbs sampler. Assessing Markov chain convergence.
12. Examples of applications of MCMC method in Bayesian statistics in SAS system. Bayesian analysis of linear regression models. Bayesian analysis of multivariate regression models. Bayesian analysis of generalized linear models. Other Bayesian models.
13. Stages of a model building –Bayesian approach vs. traditional approach. The selection of prior distributions. The selection of realizations of the Markov chain. Interpretation of results.
14. Data quality categories: intrinsic (accuracy, objectivity, believability), accessibility, contextual (relevancy, timeliness, completeness), representational (easy of understanding, concise and consistent representation).
15. Total Data Quality Management – principles, measures of data quality, and data quality improvement tools.
16. Computing values of data quality measures in various categories using real world databases, quality improvement of survey data.
17. Data normalization and standardization, variable transformation and modification.
18. Using data quality improvement tools, improving consistency and accuracy of data, data parsing.

19. Basic concepts in the analysis of missing data. Monotone and general pattern of missing data. Mechanisms that lead to missing data (MCAR, MAR and MNAR).
20. Introduction to the imputation methods. Traditional method: complete-case analysis, single imputation methods.
21. Modern data imputation methods. Likelihood-based approach. Multiple imputation.

Literatura podstawowa:

1. Frątczak E. (red.), Zaawansowane metody analiz statystycznych. Teoria – przykłady zastosowań, Oficyna wydawnicza SGH, 2012.
2. W. Grzenda. Wstęp do statystyki bayesowskiej. Oficyna Wydawnicza SGH, 2012.
3. A.Ptak-Chmielewska – Uogólnione modele liniowe , Oficyna Wydawnicza SGH, 2013.
4. Bernardo J. M., Smith A. F. M. Bayesian Theory. Wiley Series in Probability and Statistics, 2004.
5. Biemer P. P., Lyberg L. E., Introduction to Survey Quality, Wiley, New York, 2003.
6. Bolstad W. M., Introduction to Bayesian statistics. A John Wiley & Sons, 2007.
7. Gamerman D., Lopes H. F. Markov Chain Monte Carlo. Stochastic Simulation for Bayesian Inference. Second edition. Chapman & Hall (CRC Press), 2006.
8. Gelman A., Carlin J. B., Stern H. S., Rubin D. B., Bayesian data analysis. Chapman & Hall (CRC Press), 2000.
9. Hagenaars J. A., McCutcheon A. L. Applied Latent Class Analysis. Cambridge University Press, 2009.
10. J. A. Little, D. Rubin (2002) Statistical Analysis with Missing Data (John Wiley & Sons: Hoboken).
11. Lyberg L., Biemer P., Collins M., De Leeuw E., Dippo C., Schwarz N. and Trewin D. (eds), Survey Measurement and Process Quality, Wiley, New York, 1997.
12. Lynch S. M., Introduction to applied Bayesian statistics and estimation for social scientists. Springer, 2007.
13. Richard Y. Wang R. Y., Pierce E. M., Stuart E. Madnick S. E., and Fisher C. W., Information Quality, Armonk, NY: M.E. Sharpe, 2005
14. Robert Ch. P., Casella G. Monte Carlo Statistical Methods. Second Edition. Springer Texts in Statistics, 2004.
15. D. B. Rubin (1987) Multiple Imputation for Nonresponse in Surveys (John Wiley & Sons: Hoboken).
16. J. L. Schafer (1997) Analysis of Multivariate Incomplete Data (Chapman & Hall: London).
17. SAS Institute Inc., SAS/STAT SAS Online Doc, SAS Institute Inc.

Literatura uzupełniająca:

1. Agresti A. An introduction to categorical data analysis. Second edition. Wiley, 2007.
2. Korczyński A., "Review of methods for data sets with missing values and practical applications", Śląski Przegląd Statystyczny, 2014.
3. Korczyński A., "Własności estymatorów – porównanie kalibracji i imputacji wielokrotnej", Statystyka - zastosowania biznesowe i społeczne, 2014.

Część D

Prerekwizyt *(jeśli wymagany, to nazwa przedmiotu lub rodzaj wiedzy z zakresu ...)***:**

Proponowane usytuowanie przedmiotu w planie studiów:
Rok studiów:
Semestr:

Proponowana liczba punktów ECTS za przedmiot *(w stosunku do 30 ECTS za semestr)***:**

Wymiar i forma zajęć (w godzinach)			**Metody zajęć:** *(opis w jęz. polskim)* Podstawowe metody prowadzenia zajęć to wykład i zajęcia praktyczne w postaci ćwiczeń. Wykład jest prowadzony metodą tradycyjną przy zastosowaniu środków audiowizualnych. Na ćwiczeniach studenci uczą się stosować wiedzę teoretyczną do rozwiązywania konkretnych problemów i zadań przy użyciu wybranych pakietów komputerowych. Studenci opracowują własne projekty i prezentują ich wyniki. *(opis w jęz. angielskim)* The course is composed of lectures and exercises i.e. classical lecture with the use of audio-visual techniques and classes based on problem solving and practical standard exercises in software systems. The students elaborate own projects and present their results
Ogółem	Studia stacjonarne i popołudniowe	Propozycja dla studiów niestacj. sob-niedz.	
	60		Kejsy (Tak / Nie)
Wykład	32		Gry (Tak / Nie)
Ćwiczenia	28		Referaty (Tak / Nie)
Konwersatorium			Dyskusje (Tak / Nie)
Laboratorium			Przy udziale praktyków (Tak / Nie)
Inna forma *(jaka?)*			Inne *(jakie?)*
Elementy oceny końcowej (ogółem 100%), w tym:			**Charakterystyka wymagań w trakcie zajęć i na egzaminie końcowym:**
Egzamin pisemny-tradycyjny		40 %	*opis w jęz. polskim)* *Studenci w trakcie zajęć przygotowują*
Egzamin testowy			

Egzamin ustny		*projekty i je prezentują na ćwiczeniach. Egzamin jest pisemny o charakterze teoretyczno-praktycznym.* *(opis w jęz. angielskim)* *Students in the class are preparing projects and present them .* *The exam has a written form (range theoretical-practical.)*
Kolokwium		
Prace domowe, referaty, prezentacje	60%	
Ocena z ćwiczeń		
Inne *(jakie?)*		

Kryteria selekcji na zajęcia:		**Inne uwagi:**	
Lista rankingowa (Tak / Nie)		Wielkość grupy	20
Kolejność zgłoszeń (Tak / Nie)		Wymóg laboratorium komputerowego	TAK
Ocena z prerekwizytu (Tak, *jakiego?*	*/ Nie)*	Sala wyposażona w video	
Znajomość języka (Tak, *jakiego?*	*/ Nie)*	Inne *(jakie?)* *opis w jęz. polskim)* *(opis w jęz. angielskim)*	