Программа учебной дисциплины
«Анализ данных на платформе SAS»

Утверждена
Академическим советом ООП
Протокол № от «__»_____20__ г.

Автор	**Ильвовский Дмитрий Алексеевич**
Число кредитов	**3**
Контактная работа (час.)	**38**
Самостоятельная работа (час.)	**114**
Курс	**2, 3,4**
Формат изучения дисциплины	**Full time**

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Данная дисциплина ставит своей целью изучение базовых сведений по анализу данных в среде SAS. Эти знания и навыки необходимы в профессиональной деятельности специалистов по математическому моделированию и информатике.

В результате изучения дисциплины студенты должны:

- Знать основы языка SAS Base и уметь записывать и понимать простые программы на этом языке;
- Владеть основами макропрограммирования на языке SAS Base;
- Понимать принципы работы основных статистических методов анализа данных на платформе SAS;
- Уметь запускать и анализировать результаты выполнения основных статистических методов анализа данных на платформе SAS;
- Знать список основных методов анализа данных, реализованных на платформе SAS.

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

№	Наименование разделов и дисциплин	Всего час.	В том числе		Форма контроля
			лекции	Практические занятия	
1	**Раздел 1. Аналитическая платформа SAS. Обзор технологий.**		*2*	-	-

2	**Раздел 2. Язык программирования SAS/BASE**		*6*	*6*	Практические задания
2.1	Тема 2.1. Основы программирования на SAS/BASE		*4*	*4*	
2.2	Тема 2.2. Макросы, SQL		*2*	*2*	
3	**Раздел 3. Библиотека методов стат. Анализа SAS/STAT**		*10*	*8*	Практические задания
3.1	Тема 3.1. введение в SAS/STAT, дисперсионный анализ		*2*	*2*	
3.2	Тема 3.2. Линейная регрессия		*2*	*2*	
3.3	Тема 3.3.Логистическая регрессия		*2*	*2*	
3.4	Тема 3.4. Обобщенные линейные модели		*2*	*2*	
			18	*14*	
Итоговый контроль			***Экзамен***		

III. ОЦЕНИВАНИЕ

Дисциплина «Анализ данных на платформе SAS» читается в 3 модуле.

Тип контроля	Форма контроля	Параметры
Текущий контроль	Домашнее задание	Выдается для поэтапного выполнения в течение модуля
Итоговый контроль	Экзамен	Письменная работа 80 минут

На текущем и итоговом контроле студент должен продемонстрировать владение основными понятиями из пройденных тем дисциплины.

Текущий контроль включает письменное задание, состоящее из нескольких задач по пройденному материалу.

Итоговый контроль проводится в форме письменного экзамена, включающего несколько вопросов и задач по темам дисциплины.

Порядок формирования оценок по дисциплине

Преподаватель оценивает самостоятельную работу студентов по выполнению домашних работ, выдаваемых на семинарских и практических занятиях. При этом оценивается правильность, эффективность и оформление программного кода. Оценки за домашние задания выставляются в рабочую ведомость, и перед экзаменом в конце 3-го модуля за домашние задания выставляется результирующая оценка по десятибалльной шкале $O_{сам.\ работа}$. Оценка за домашнее задание, сданное позднее объявленного срока, понижается на 2 балла (но не ниже 5 баллов).

Оценка итогового контроля в конце 3-го модуля выставляется по следующей формуле:

$$O_{итог} = 0{,}5 \cdot O_{экзамен} + 0{,}5 \cdot O_{сам.\ работа}.$$

и округляется до целого числа арифметическим способом,

где $O_{экзамен}$ – оценка за работу непосредственно на экзамене по десятибалльной системе.

В случае пропусков занятий и домашних заданий студент может сдать все домашние задания не позднее чем за 5 дней до экзамена.

Студенты, имеющие $O_{сам.\ работа} = 10$, автоматически получают 10 баллов за экзамен после прохождения короткого устного собеседования с преподавателем.

Сертификат вручается студентам, имеющим итоговую оценку не ниже 9 баллов.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

1. Используя процедуру MEANS, познакомиться с данными. [P1] Вывести средние в новый набор данных. [P2] Построить график процедурой SGPLOT, используя полученный набор данных: strength но оси Y, Additive по оси X, группировать по переменной Brand. [P3] Что вы можете сказать о данных? [P4] Основываясь на графике, нужно ли использовать пересечение факторов Additive и Brand в модели?
2. [P1] Проверьте гипотезу о том, что средняя прочность одинакова для всех марок. Проверить предположения. Если возможно, сравните все марки с маркой Graystone. [P2] Добавьте оставшийся фактор – Additive. Какие выводы можно сделать сейчас? [P3] Если графики из п.1 говорят, что нужно использовать пересечение, то добавьте его. Какие выводы вы можете сделать на данном шаге анализа?

3. Выполните подходящие множественные сравнения для статистически значимых переменных.

Демонстрационный вариант экзамена

1. В исходной таблице TAB1 есть (в случайном порядке) 95% данных с входной переменной X и с целевой переменной Y, равной 1, и 5% данных с её значением, равным нулю. Создайте такую обучающую выборку TAB2 , куда попадут все наблюдения с целевой переменной, равной нулю, и столько же наблюдений с целевой переменной, равной единице (строки с единицами можно выбрать как угодно: последовательно, случайно, каждую N-ую и т. д.)
2. Пусть есть таблица work.tab1, содержащая (среди прочих) переменную a (числовую, интервальную). Напишите шаг данных для стандартизации значений в переменной a (Z-scoring), то есть их замене по формуле a=(a-mu)/sigma, где mu - выборочное мат. ожидание переменной a, a sigma - среднеквадратичное отклонение переменной a. При вычислении значений mu и sigma игнорируйте пропущенные значения.
3. Дан код:
   %let a=10; data test; r= 0; run ;

   Какие макроинструкции вам обязательно понадобятся, чтобы вывести "a" раз наблюдение r=0 в набор данных test :

   (%for) (for)(%macro) (output) (%output) (%do) (do) (end) (%end) (%mend)

4. Выбирая значимые переменные с помощью процедуры PROC VARCLUS, ориентируются на критерий 1-R^2, потому что он позволяет:

a) выбрать предикторы, сильнее всего связанные с целевой переменной
b) оценить оптимальное количество кластеров
c) сгруппировать в один кластер несколько коррелирующих переменных
d) выбрать наиболее репрезентативный предиктор из группы коррелирующих признаков

V. РЕСУРСЫ

5.1 Базовая литература – документация SAS (http://support.sas.com/documentation):

1. SAS/STAT(R) 9.4 User's Guide
2. SAS(R) 9.4 Functions and CALL Routines: Reference

Основная литература – online-курсы обучения:

3. Base SAS(R) 9.4 Procedures Guide, Second Edition
4. SAS(R) 9.4 SQL Procedure User's Guide.

5.2 Программное обеспечение

Программная среда для работы:

http://www.sas.com/en_us/software/university-edition.html

5.3 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

– ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);

– мультимедийный проектор с дистанционным управлением.

Учебные аудитории для лабораторных и самостоятельных занятий по дисциплине оснащены компьютерами, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.

Методические указания по освоению дисциплины.

Для лучшего усвоения дисциплины рекомендуется:

- пройти онлайн-курсы:
    - SAS Programming I: Essentials https://support.sas.com/edu/schedules.html?ctry=us&id=277
    - SAS Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression https://support.sas.com/edu/schedules.html?ctry=us&id=1979
- читать:
    - Г.И.Ивченко, Ю.И.Медведев. Математическая статистика
    - О дисперсионном анализе в SAS/STAT:

http://support.sas.com/documentation/cdl/en/statug/66103/HTML/default/viewer.htm#statug_introanova_toc.htm