Министерство образования Республики Беларусь

Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет Информационных технологий и управления
Кафедра Интеллектуальных информационных технологий

Индивидуальная практическая работа №4
по дисциплине «Статистические основы индуктивного вывода»
на тему
Регрессионный анализ в MS Excel

Студент гр. 021703: Е. С. Колосовский

Проверил: А. А. Ефремов

Минск 2023

1 Предсказывание стоимости страховки на основе регрессионного анализа

1.1 Задачи

а Привести таблицу с исходными данными.

б Описать, что анализируется: какой показатель и в зависимости от каких показателей.

в Рассказать о проверке выборки на наличие аномальных наблюдений по критерию Граббса.

г Привести корреляционную матрицу. Сделать вывод о наличии/отсутствии мультиколлинеарности.

д Привести протокол первичного регрессионного анализа из Excel.

е Сделать вывод о статистической значимости коэффициентов уравнения по Стьюденту.

ж Сделать вывод о качестве построенной модели по критерию Фишера.

з Рассчитать среднюю ошибку аппрокимации.

и Расписать подробно экономическую интерпретацию каждого (значимого) коэффициента полученного конечного уравнения регрессии.

к Выполнить точечный и интервальный прогноз. Сравнить этот прогноз с фактическими данными одного дополнительного наблюдения, не вошедшего в начальную выборку. Вычислить относительную погрешность прогноза (в %).

л Выбрать наиболее влиятельный из факторных показателей и построить по нему модель парной нелинейной регрессии, используя все известные типы тренда, включая гиперболический и обратный линейный.

м Выбрать наилучшее уравнение по критериям Фишера и Стьюдента. Привести соответствующий этому уравнению график.

н Привести проверку остатков на гетероскедастичность.

о Выполнить прогноз, используя наблюдение, о котором шла речь в пункте 10. Вычислить относительную погрешность прогноза.

п Сделать краткий вывод по работе в целом и по возможности дальнейшего практического использования результатов.

1.2 Исходные данные

В качестве датасета будет использоваться: `https://www.kaggle.com/datasets/mirichoi0218/insurance`

Известна следующая информация:

1. Возраст
2. Пол
3. Индекс массы тела
4. Количество детей
5. Курильщик
6. Регион проживания
7. Стоимость страховки — y

Часть датасета:

Возраст	Пол	ИМТ	Детей	Курит	Регион	Стоимость страховки
19	female	27.9	0	yes	southwest	16884.924
18	male	33.77	1	no	southeast	1725.5523
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.47061
32	male	28.88	0	no	northwest	3866.8552
31	female	25.74	0	no	southeast	3756.6216
46	female	33.44	1	no	southeast	8240.5896
37	female	27.74	3	no	northwest	7281.5056
37	male	29.83	2	no	northeast	6406.4107
60	female	25.84	0	no	southwest	28823.13692
25	male	26.22	0	no	southeast	2721.3208
62	female	26.29	0	yes	southeast	27808.7251
23	male	34.4	0	no	southwest	1826.483
56	female	39.82	0	no	southeast	11090.7178
27	male	42.13	0	yes	southeast	39611.7577
19	male	24.6	1	no	southwest	1837.237
52	female	30.78	1	no	northeast	10797.3362
23	male	23.845	0	no	northeast	2395.17155
56	male	40.3	0	no	southwest	10602.385
30	male	35.3	0	yes	southwest	36837.467
60	female	36.005	0	no	northeast	13228.84695
30	female	32.4	1	no	southwest	4149.736
18	male	34.1	0	no	southeast	1137.011
34	female	31.92	1	yes	northeast	37701.8768
37	male	28.025	2	no	northwest	6203.90175
59	female	27.72	3	no	southeast	14001.1338
63	female	23.085	0	no	northeast	14451.83515
55	female	32.775	2	no	northwest	12268.63225
23	male	17.385	1	no	northwest	2775.19215

Таблица 1 – Стоимость страховки.

1.3 Корреляционная матрица

Correlations (Insurance)
Marked correlations are significant at p < ,05000
N=545 (Casewise deletion of missing data)

Variable	age	sex	bmi	smoker	children	charges
age	1,000000	0,535997	0,366494	0,517545	0,420712	0,384394
sex	0,535997	1,000000	0,151858	0,193820	0,083996	0,352980
bmi	0,366494	0,151858	1,000000	0,373930	0,408564	0,139270
smoker	0,517545	0,193820	0,373930	1,000000	0,326165	0,177496
children	0,420712	0,083996	0,408564	0,326165	1,000000	0,045547
charges	0,384394	0,352980	0,139270	0,177496	0,045547	1,000000

Рис. 1 – Корреляционная матрица

Из полученной матрицы видно, что достаточное влияние на стоимость страховки имеют параметры: возраст, пол, курит человек или нет, ИМТ. Мультиколлинеарность отсутствует,

так как все параметры связаны между собой недостаточно сильно.

1.4 Анализ на выбросы с помощью критерия Граббса

Было решено оценивать выбросы по наиболее значимому параметру - возраст. 1

Датасет был отсортирован по увеличению возраста. Затем были найдены значения разностей:

$$\delta_i = x_{i+1} - x_i$$

После этого было найдено среднее значение разностей:

$$\delta' = \frac{1}{n'} \sum_{i=1}^{n} n\delta_i, n' = n - i$$

Было рассчитано среднеквадратичное отклонение и статистика Граббса соответственно:

$$S^2 = \frac{1}{n' - 1} \sum_{i}^{n} n(\delta_i - \delta')^2$$

$$G_{i'} = \frac{|\delta' - \delta_{i'}|}{S}$$

Она вышла равной 0,053. По таблице Граббса было найдено критическое значение, равное 3,107. Т.к. $G_i < G$, то выбросов в датасете нет.

1.5 Первичный регрессионный анализ

ВЫВОД ИТОГОВ					Y-пересечение	Deep sleep percen	Awakenings	
					0,584816575	0,004799752	-0,029431615	
Регрессионная статистика								
Множественный I	0,822750621							
R-квадрат	0,676918585							
Нормированный	0,662559411							
Стандартная оши	0,067529507							
Наблюдения	48							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*Значимость F*			
Регрессия	2	0,429956121	0,214978061	47,1418889	9,11002E-12			
Остаток	45	0,205210545	0,004560234					
Итого	47	0,635166667						
	Коэффициенты	*Стандартная ошиб.*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	0,584816575	0,038820188	15,06475393	3,40235E-19	0,506628704	0,663004447	0,506628704	0,663004447
Deep sleep percen	0,004799752	0,000615089	7,803346817	6,67302E-10	0,0035609	0,006038605	0,0035609	0,006038605
Awakenings	-0,029431615	0,007105208	-4,142259493	0,000149402	-0,043742238	-0,015120991	-0,043742238	-0,015120991

1.6 Ошибка аппроксимации

Наблюдение	казанное Sleep effi	Остатки	процент ошибки
1	0,88720098	0,03279902	3,696909861
2	0,891367632	0,058632368	6,57779865
3	0,57268467	0,02731533	4,769698075
4	0,819371345	0,120628655	14,72209829
5	0,675378772	-0,125378772	18,56421567
6	0,59251678	0,03748322	6,326102677

Суммарная средняя ошибка составила 7,062920945%. Модель является достаточно хорошей.

1.7 Точечный и интервальный прогноз

Интервал составил 0,143233.

Получается, что допустимый диапазон 0,54 ± 0,14. Значение 0,6459 попадает в него.

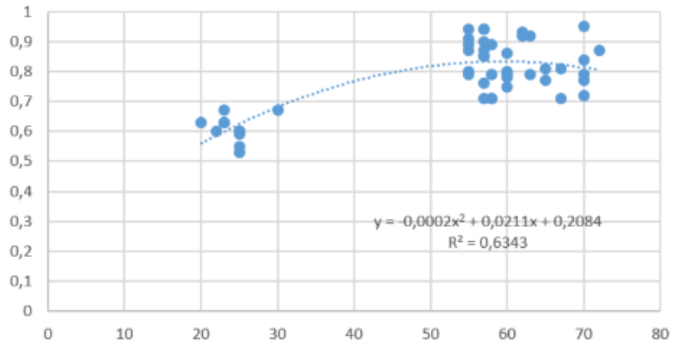
1.8 Построение нелинейной регресии

Наиболее значимым признаком является возраст.

Рассмотрим различные виды линии тренда. В таблице ниже представлен вид линии тренда и его показатель R^2.

линейн	0,5537
лог	0,5904
пол	
2	0,6343
степ	0,6342

1.9 Полиномиальное уравнение 2-ой степени



Статистические переменные модели можно назвать удовлетворительными.

1.10 Проверка на гетероскедастичность

Разделим выборку на 3 группы и проверим 1-ую и 3-ую на гетероскедастичность.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный	0,936575739
R-квадрат	0,877174116
Нормированный	0,858277826
Стандартная оши	0,051816277
Наблюдения	16

Дисперсионный анализ

	df	*SS*	*MS*	*F*
Регрессия	2	0,249270954	0,124635477	46,42044122
Остаток	13	**0,034904046**	0,002684927	
Итого	15	0,284175		

	Коэффициенты	*андартная ошиб*	*t-статистика*	*P-Значение*
Y-пересечение	0,371651782	0,053032067	7,008057606	9,2295E-06
Deep sleep percer	0,008961199	0,0010066	8,902446088	6,81093E-07
Awakenings	0,008487849	0,011614673	0,730786753	0,477874494

Рис. 2 – 1-ая группа

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный F	0,756846829
R-квадрат	0,572817123
Нормированный I	0,50709668
Стандартная оши	0,05339893
Наблюдения	16

Дисперсионный анализ

	df	*SS*	*MS*	*F*
Регрессия	2	0,049706206	0,024853103	8,715965681
Остаток	13	**0,037068794**	0,002851446	
Итого	15	0,086775		

	Коэффициенты	*андартная ошиб*	*t-статистика*	*P-Значение*
Y-пересечение	1,089079032	0,230943512	4,715781032	0,000403559
Переменная X 1	-0,002840009	0,003450944	-0,822965687	0,425366815
Переменная X 2	-0,035728207	0,008596334	-4,156214224	0,00112805

Рис. 3 – 2-ая группа

Расчеты представлены ниже:

F_расч	1,06202
F_кр	3,80557

Так как рассчетное значение меньше критического, то остатки гомоскедатичные.

1.11 Прогноз по нелинейной регрессии

Возьмем полиномиальную модель.

57	3249	0,76		предсказ	ошибка
25	625	0,54		0,6109	13,12963

Рассчитаем ошибку по формуле:

$$S_{\hat{y}} = S * \sqrt{1 + \frac{1}{n} + \frac{\left(x_{пр} - \bar{x}\right)^2}{n * \sigma_x^2}}$$

Получим:

ошибка	0,100497
t_кр	2,014103
интервал	0,202412

Предсказание вписывается в интервал.

1.12 Вывод:

Была построена модель, которая предсказывает стоимость страховки для людей в зависимости от различных характеристик. Данный анализ можно использовать для прогнозирования стоимости страхования жизни человека, исходя из того как он ведет свою жизнь, его возраста и других критериев.