**Федеральное агентство по образованию Российской Федерации**

**Государственное образовательное учреждение**

**высшего профессионального образования**

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**Кафедра анализа данных и искусственного интеллекта**

Преподаватель,

д.т.н. А.А. Халафян

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ОТЧЕТ**

**СТАТИСТИЧЕСКИЙ АНАЛИЗ КРУПНЕЙШИХ СТРАХОВЩИКОВ ИМУЩЕСТВА В 2007Г**

Работу выполнил студент 4 курса

…………………………………………..

……………………………………………….

**Отсутствует Подгонка распределений, Непараметрическая статистика!!! Описаны методы, по которым не были прочитаны лекции, в описании результатов других методов очень много ошибок, переделать отчет**

Краснодар 2024

СОДЕРЖАНИЕ

[**1 Исходные данные 3**](#_Toc233138293)

[**2 Графический анализ 5**](#_Toc233138294)

[**2.1 2D Graphs 5**](#_Toc233138295)

[**2.1.1 2D Histogramms 5**](#_Toc233138296)

[**2.1.2 2D Scatterplots 6**](#_Toc233138297)

[**2.1.3 2D Box Plots 9**](#_Toc233138298)

[**2.2 Средство «закрашивание» 10**](#_Toc233138299)

[**2.3 3D SequentialGraphs 12**](#_Toc233138300)

[**2.3.1 Raw Data Plots 12**](#_Toc233138301)

[**2.3.2 Bivariate Histograms 13**](#_Toc233138302)

[**2.3.3 Box Plots 13**](#_Toc233138303)

[**3 Основные статистики 15**](#_Toc233138304)

[**3.1 Описательные статистики 15**](#_Toc233138305)

[**3.2 Корреляционная матрица 15**](#_Toc233138306)

[**3.3 Критерий Стьюдента сравнения средних 16**](#_Toc233138307)

[**3.3.1 t-test, independent, by groups 16**](#_Toc233138308)

[**3.3.2 t-test, independent, by variables 16**](#_Toc233138309)

[**3.3.3 t-test, dependent samples 16**](#_Toc233138310)

[**3.3.4 t-test, single samples 17**](#_Toc233138311)

[**3.4 Группировка и однофакторная ANOVA 17**](#_Toc233138312)

[**4 Частотный анализ 19**](#_Toc233138313)

[**4.1 Таблицы частот 19**](#_Toc233138314)

[**4.2 Таблицы кросстабуляции 21**](#_Toc233138315)

[**5 Корреляционный анализ 23**](#_Toc233138316)

[**6 Дисперсионный анализ 24**](#_Toc233138317)

[**7 Линейное многомерное моделирование взаимосвязей 27**](#_Toc233138318)

[**8 Нелинейное многомерное моделирование взаимосвязей 29**](#_Toc233138319)

[**8.1 Fixed Nonlinear Regression 29**](#_Toc233138320)

[**8.2 Логит регрессия 31**](#_Toc233138321)

[**8.3 ~~Пробит регрессия~~ 32**](#_Toc233138322)

[**8.4 ~~Экспоненциальная регрессия~~ 33**](#_Toc233138323)

[**8.5 ~~Кусочно-линейная регрессия~~ 34**](#_Toc233138324)

[**~~8.6 Определенная пользователем регрессия~~ 35**](#_Toc233138325)

[**~~9 Канонический анализ 36~~**](#_Toc233138326)

[**10 Дискриминантный анализ 41**](#_Toc233138327)

[**11 Классификационный анализ без обучения 48**](#_Toc233138328)

[**11.1 Кластерный анализ 48**](#_Toc233138329)

[**11.1.1 Метод к-средних 48**](#_Toc233138330)

[**11.1.2 Двухвходовая кластеризация 50**](#_Toc233138331)

[**11.1.3 Древовидная кластеризация 52**](#_Toc233138332)

[**11.2 Деревья классификации 53**](#_Toc233138333)

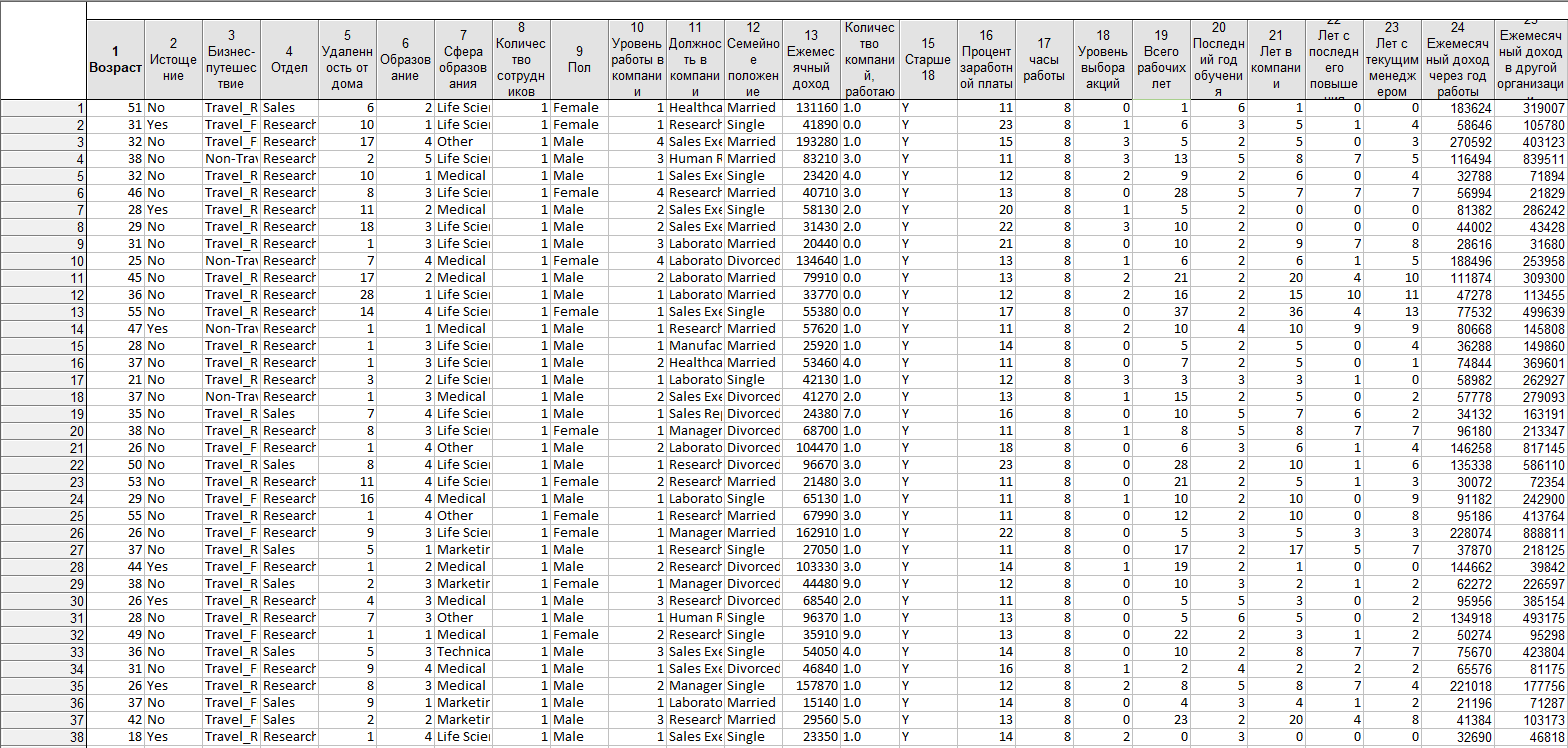
[**12 Методы редукции данных 57**](#_Toc233138334)

[**12.1 Факторный анализ 57**](#_Toc233138335)

[**~~12.2 Метод анализ главных компонент и классификация 61~~**](#_Toc233138336)

[**~~13 Многомерное шкалирование 65~~**](#_Toc233138337)

**1 Исходные данные**



В исходной таблице представлены данные о сотруднике и его работе.

Показатели:

1. Возраст – возраст сотрудника
2. Истощение – уволился ли сотрудник в предыдущем году
3. Бизнес-путешествие – как часто сотрудники совершали командировки в деловых целях за последний год
4. Отдел – отдел в компании
5. Удаленность от дома – расстояние от дома в км
6. Образование – уровень образования в шкале от 1 до 5
7. Сфера образования – сфера образования сотрудника
8. Количество сотрудников – количество сотрудников
9. Пол – пол сотрудника
10. Уровень работы в компании – уровень работы в шкале от 1 до 5
11. Должность в компании – должность сотрудника
12. Семейное положение – семейное положение сотрудника
13. Ежемесячный доход – доход сотрудника в рублях в месяц
14. Количество компаний – количество компаний, в которых работал сотрудник
15. Старше 18 – сотрудник старше 18 или нет
16. Процент заработной платы – процентное повышение заработной платы за прошлый год
17. Часы работы – часы работы сотрудника
18. Уровень выбора акций – уровень опционов на акции сотрудника
19. Всего рабочих лет – стаж сотрудника
20. Число обучений за прошлый год – количество раз обучения сотрудника в прошлом году
21. Лет в компании – сколько лет сотрудник в компании
22. Лет с последнего повышения – Количество лет с момента последнего повышения
23. Лет с текущим менеджером – сколько лет работает под руководством текущего менеджера
24. Ежемесячный доход через год работы – ежемесячный доход спустя год работы в компании
25. Ежемесячный доход в другой организации – доход сотрудника в другой организации

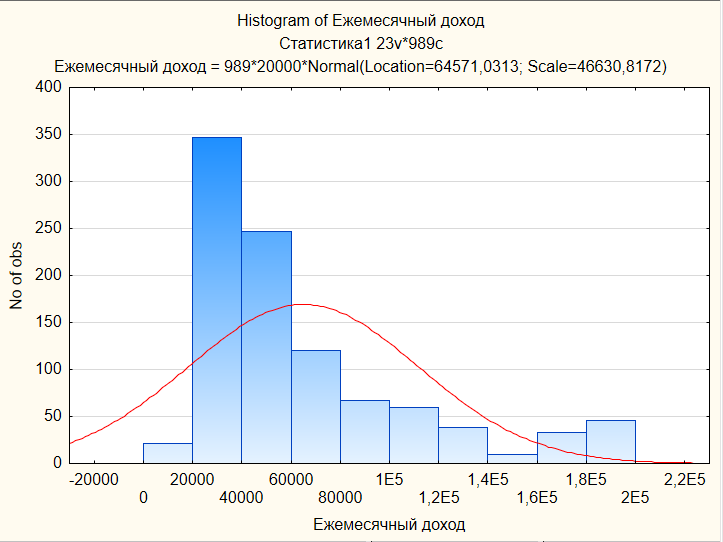
**2 Графический анализ**

**2.1 2D Graphs**

**2.1.1 2D Histogramms**

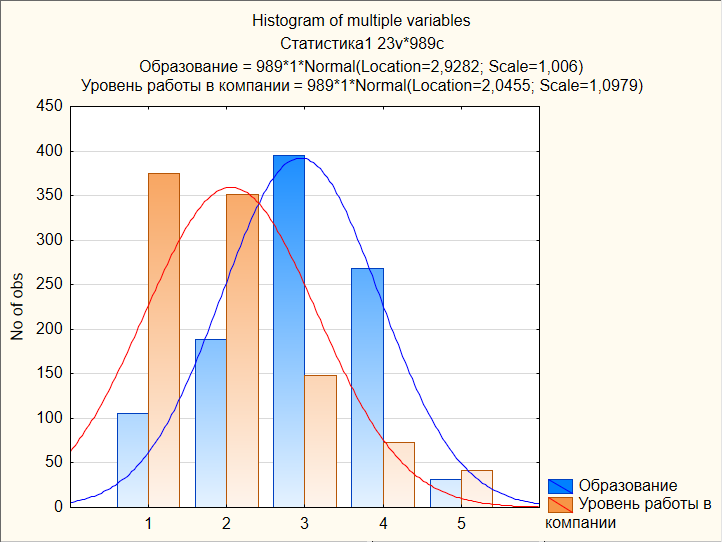
2D Histogramms являются графическими представлениями распределения частот выбранных переменных.

2D Histogramms Regular (простые) – столбчатая диаграмма распределения частот.



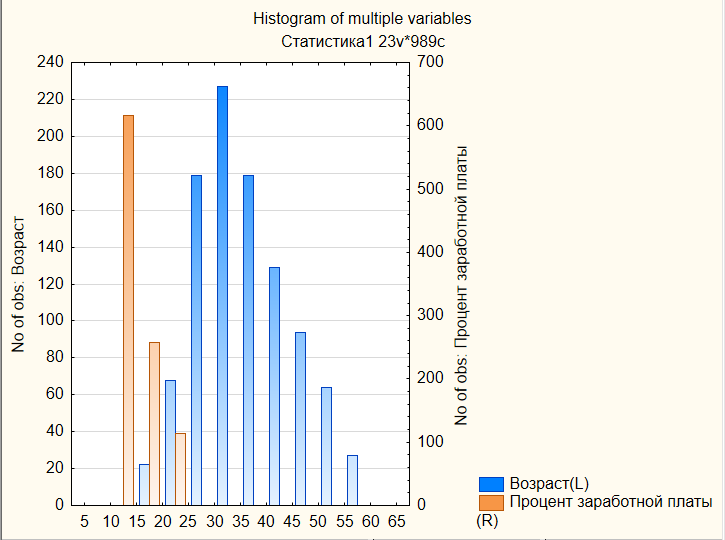
Данная 2D Histogramms Regular построена по ежемесячному доходу сотрудника в компании. Из гистограммы видно, что чаще всего в компании встречаются сотрудники с зарплатой в диапазоне от 20000 до 40000.

2D Histogramms Multiple (составные) – изображают распределение частот для нескольких переменных на одном графике.



Данная 2D Histogramms Multiple построена по уровню образования и уровню работы в компании с заданными границами: 0<=x<=1, 1<x<=2, 2<x<=3, 3<x<=4, 4<x<=5. Из гистограммы видно, что при выше заданных границах интервалов, больше всего встречаются сотрудники с уровнем образования и уровнем работы в компании в интервале [1, 3].

2D Histogramms Double-Y (с двойной осью Y) – Комбинация двух по-разному масштабированных составных гистограмм.

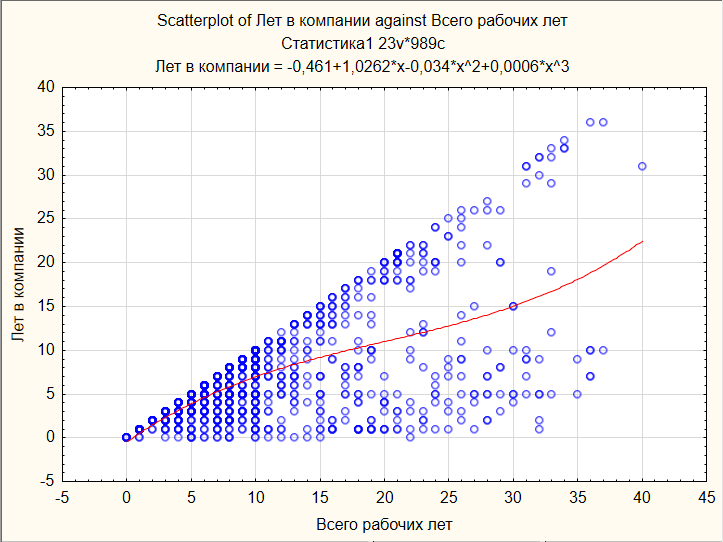


Данная 2D Histogramms Double-Y построена по двум разномасштабным переменным: возраст (лет) и процент заработной платы (процент). Из гистограммы видно, что чаще всего встречаются сотрудники с повышением зарплаты в диапазоне от 10 до 15 процентов и с возрастом от 30 – 35 лет.

**2.1.2 2D Scatterplots**

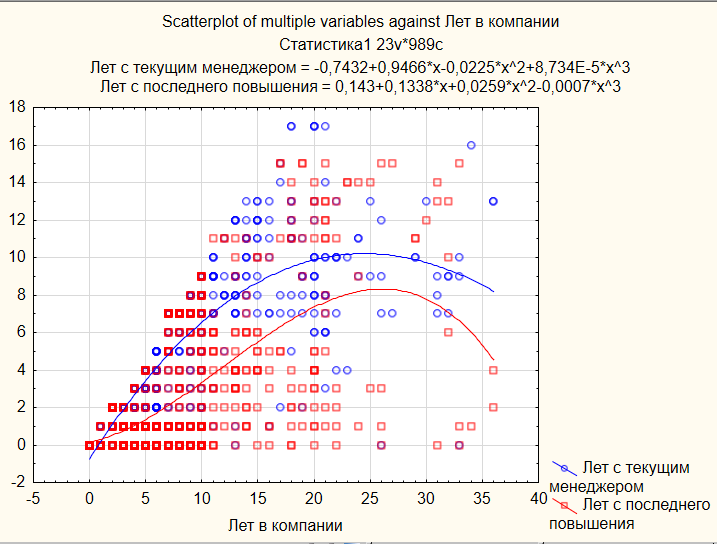
2D Scatterplots (диаграммы рассеяния) визуализируют зависимость между двумя переменными.

2D Scatterplots Regular (протые) - визуализируют зависимость между двумя переменными X и Y.



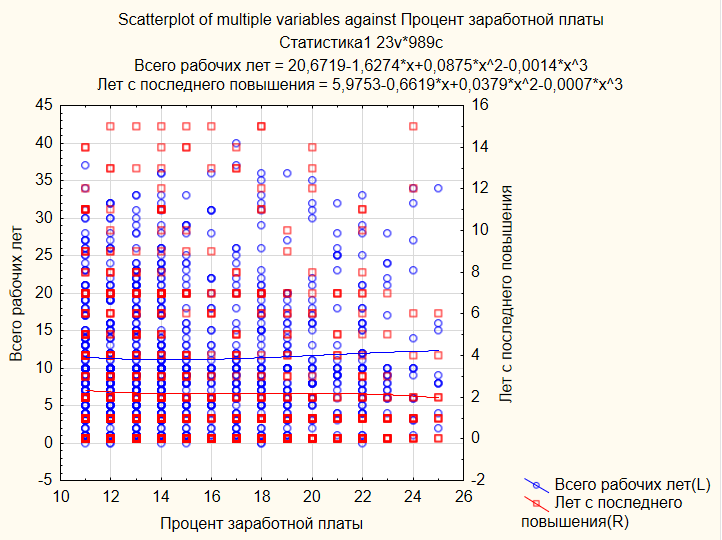
Данная 2D Scatterplots Regular диаграмма отображает зависимости между количеством отработанных лет в текущей компании и общим стажем сотрудника.

2D Scatterplots Multiple (составные) – состоит из нескольких зависимостей и изображает несколько взаимосвязей.



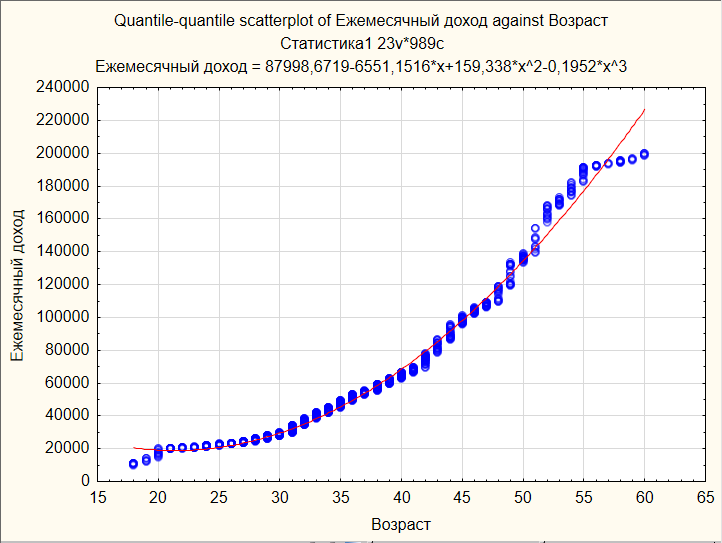
Данная 2D Scatterplots Multiple диаграмма отображает зависимости между количеством лет в текущей компании и количеством лет работы с текущим менеджером и лет с последнего повышения.

2D Scatterplots Double-Y (с двойной осью Y) – комбинация двух составных диаграмм рассеяния для одной переменной X и двух различных наборов переменных Y.



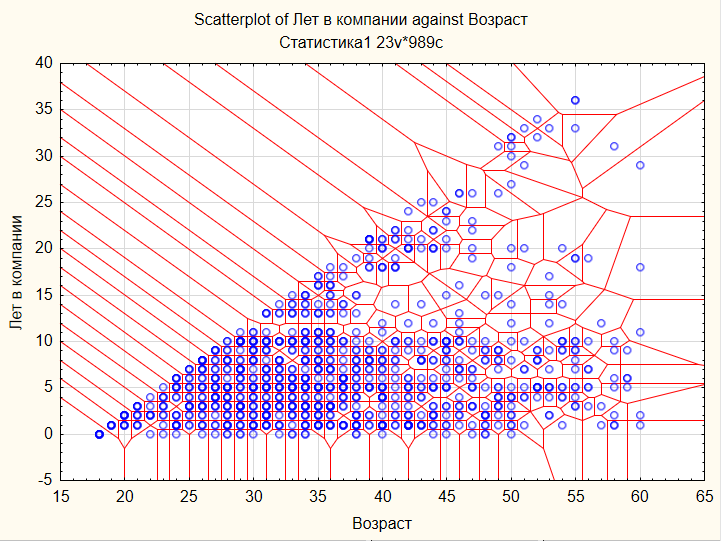
Данная 2D Scatterplots Double-Y диаграмма отображает комбинацию двух составных диаграмм: первая зависимость между процентом повышения заработной платы и стажем сотрудника, вторая зависимость между процентом повышения заработной платы и количеством лет с последнего повышения.

2D Scatterplots Quantile (квантилей) – показывает зависимость между квантилями двух переменных, позволяющая оценить сходство эмпирических распределений.



Данная 2D Scatterplots Quantile отображает зависимость между квантилями возраст и ежемесячный доход.

2D Scatterplots Voronoi (Вороного) – показывает разделение пространства между точками данных. Пространство между отдельными точками делится границами на такие области, каждая точка которых находится ближе к заключенной внутри точке, чем к любой другой соседней точке.



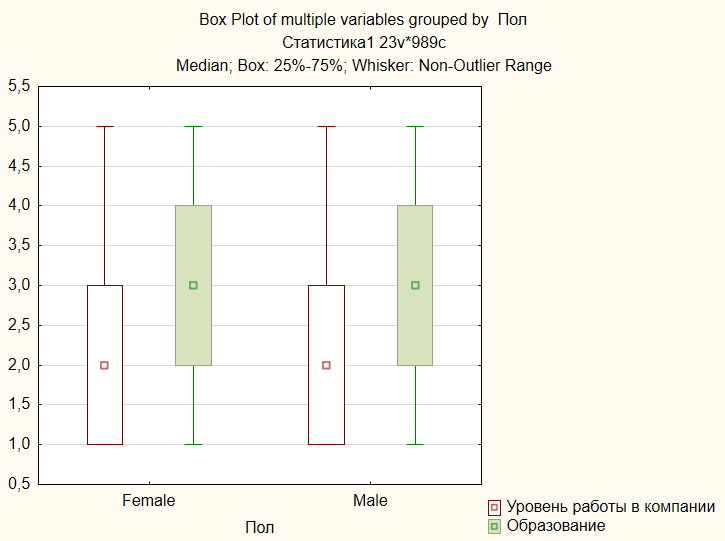
Данная 2D Scatterplots Voronoi диаграмма построена для переменных возраст и количество лет в компании.

**2.1.3 2D Box Plots**

2D Box Plots (графики ящика – диаграммы размаха) – на этих диаграммах изображаются диапазоны или характеристики распределения значений выбранной переменной отдельно по группам, заданным категориальной переменной.



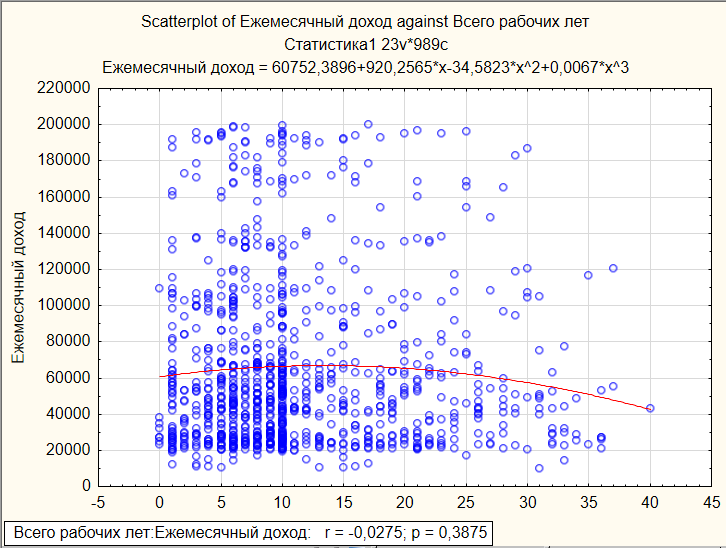
На этой 2D Box Plots Regular диаграмме представлены характеристики распределения значений переменной ежемесячный доход по группам женский и мужской пол.



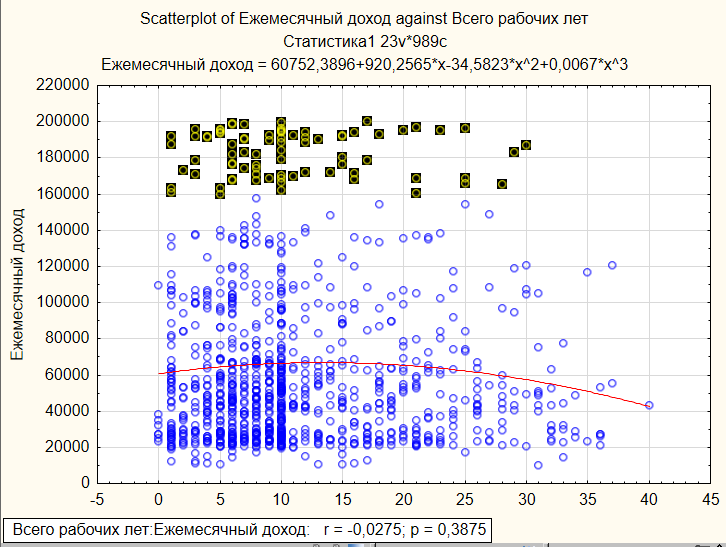
На этой 2D Box Plots Multiple диаграмме представлены характеристики распределения значений уровень работы в компании и уровень образования в зависимости от групп женский и мужской пол.

**2.2 Средство «закрашивание»**

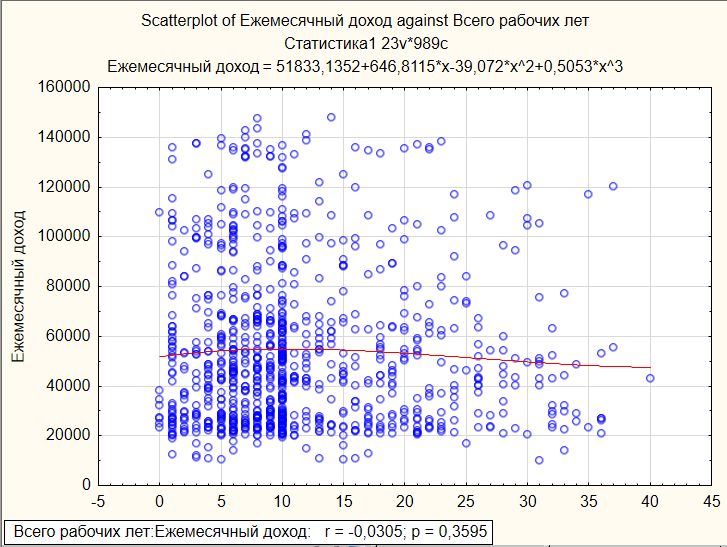
Построим диаграмму рассеивания для переменных ежемесячный доход и всего рабочих лет:



Из диаграммы видно, что есть точки, которые располагаются далеко от линии регрессии, поэтому аппроксимация может быть некачественной. Для исключения этой ситуации воспользуемся средством «закрашивание», чтобы исключить некоторые точки. Предварительно можно узнать каким страховщикам соответствуют такие точки:



Теперь с помощью тех же средств закрашивания можно исключить эти точки:

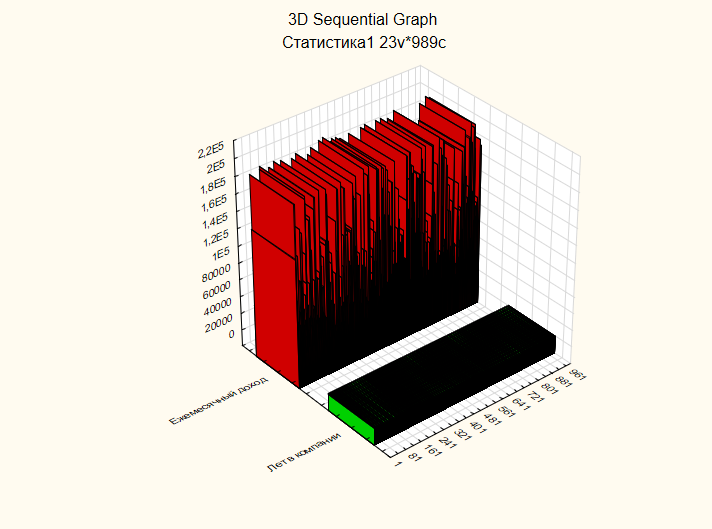


Сравнивая r до и после исключения точки, можно сказать, что качество аппроксимации увеличилось (доля объясненной дисперсии).

**2.3 3D SequentialGraphs**

**2.3.1 Raw Data Plots**

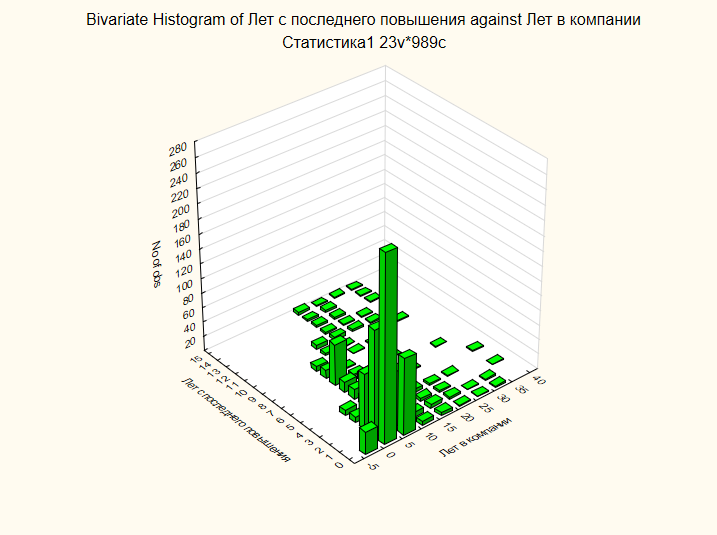
Raw Data Plots (графики исходных данных) – иллюстрируют соотношения между значениями переменных.



На данном графике показано соотношение значений между ежемесячным доходом и количеством лет в компании.

**2.3.2 Bivariate Histograms**

Эти гистограммы можно рассматривать как сочетание двух простых гистограмм, соединенных так, чтобы можно было исследовать частоты совместного появления значений двух переменных.

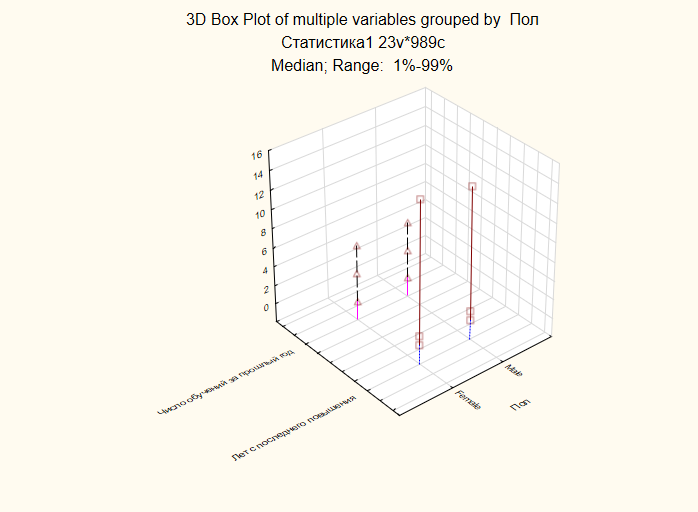


На данной гистограмме показано совместное распределение частот переменных количество лет в компании и количество лет с последнего повышения.

**2.3.3 Box Plots**

Box Plots - это диаграммы размаха.

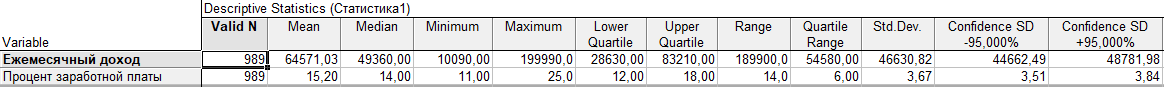
Построим диаграмму размаха для переменных число обучений за последний год и количество лет с последнего повышения по группам женский и мужской пол:



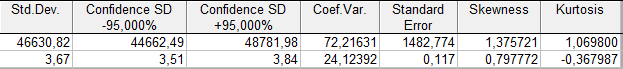
3 **Основные статистики**

**3.1 Описательные статистики**

Таблица описательных статистик для итогов по страхованию имущества:

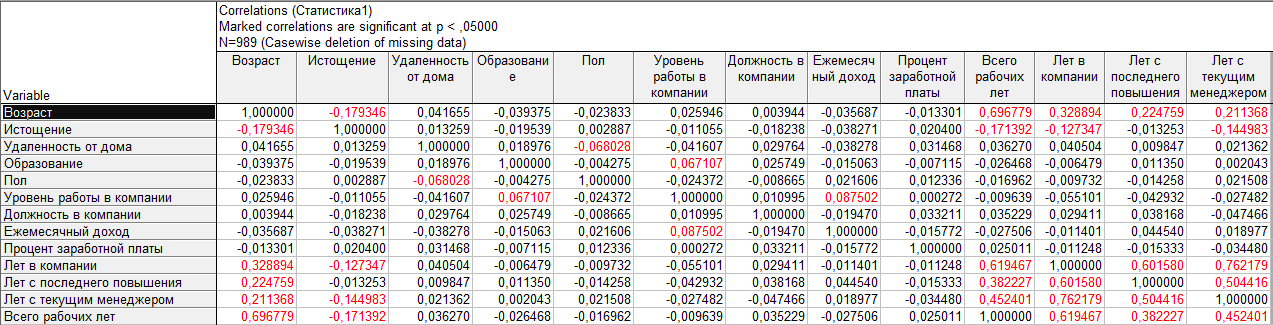
****

*продолжение*



В таблице представлены следующие статистики: число наблюдений 989, среднее 64571,03, медиана 49360,00, минимальное значение: 10090,00, максимальное значение: 199990,00, нижний квантиль: 28630,00, верхний квантиль: 83210,00, разность межу макс. и мин. значениями выборки: 189900,00, разность межу верхним и нижним значениями квантилей: 54580,00, стандартное отклонение: 46630,82, доверительный интервал стандартного отклонения -95,000%: 44662,49, доверительный интервал стандартного отклонения +95,000%: 48781,98, коэффициент вариации: 72,22, стандартная ошибка среднего: 1482,77, асимметрия: 1,38 – распределение имеет длинный правый хвост, эксцесс: 1,07 – распределение имеет заостренный пик.

**3.2 Корреляционная матрица**



Из таблицы видно, что сильная корреляция только между лет с текущим менеджером и лет в компании; умеренная между всего рабочих лет и возраст, лет в компании и всего рабочих лет, лет в компании и возраст, лет с последнего повышения и всего рабочих лет, лет с текущим менеджером и всего рабочих лет, лет в компании и лет с последнего повышения, лет с текущим менеджером и лет с последнего повышения, а слабая везде, где корреляция <=0.25. Красным отмечены значения, где уровень значимости р <0.05, т.е. статистически значимые взаимосвязи.

**t-test, independent**, **by groups** применяется если сравниваются средние количественной переменной значения которой разбиваются группирующей переменной на 2 группы!

**t-test, independent, by variablesи** применяется, если сравниваются 2 столбца одной и той же величины, но записанные для разных объектов, или для одних и тех же объектов но разные величины с одинаковыми единицами измерения, сравнение которых имеет смысл!!! У вас в данных нет таких столбцов, поэтому этот анализ либо не проводим, либо придумываем 2 столбец, например цена автомобиля, но оговариваем, что это цена другой группы авто!

**t-test, dependent** **samples** применяем, если сравнивается одна и та же величина при 2 повторных измерениях! Такой столбец надо также придумать!

**t-test, single samples** применяем, если сравниваем среднее группы со средним генеральной совокупности, которую тоже надо придумать!

**Группировка и однофакторная ANOVA** применяем, если сравниваем более 2 групп по одной, или более группирующим переменным!!!!!!!!

**Дисперсионный анализ,** если сравниваем подгруппы по **двум**  группирующим переменным

**Непараметрические критерии**

**Ранговый ДА и конкордация Кендалла** применяем, если сравниваем более 2 повторных измерений!!!!

**критериями знаков и Вилкоксона** применяем, сравнивается одна и та же величина при 2 повторных измерениях

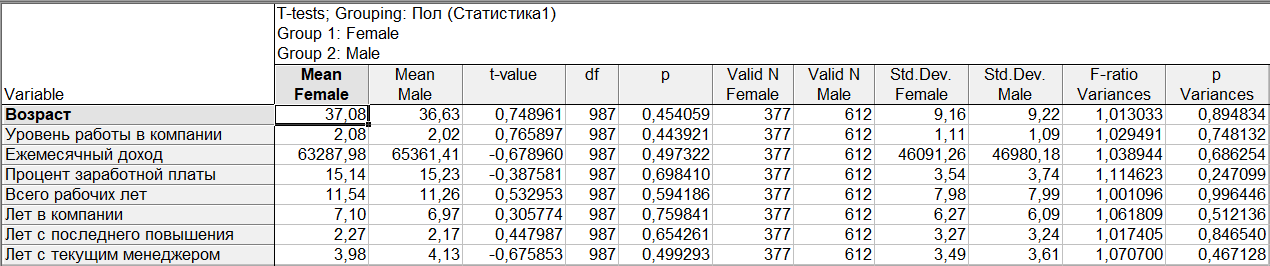
**Критерий манна-Уитни, Вальда-Вольфовица, Колмогорова-Смирнова** (у вас их нет!, надо добавить!) если сравниваются средние количественной переменной значения которой разбиваются группирующей переменной на 2 группы!

Если групп более 2, то применяем **критерий Краскела-Уоллиса**

**3.3 Критерий Стьюдента сравнения средних**

**3.3.1 t-test, independent, by groups**

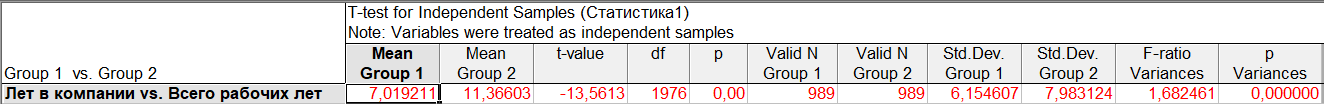
Проверим равенство средних для всех переменных по группам прибыльных и убыточных страховщиков на основе t-критерия (t-test, independent, by groups).



По данным таблицы можно сделать вывод, что средние для всех переменных отличаются статистически незначимо (верна гипотеза о равенстве средних в генеральных совокупностях), так как уровни значимости р критерия Стьюдента больше, чем 0,05, и гипотеза о равенстве дисперсий справедлива для всех переменных.

**3.3.2 t-test, independent, by variables**

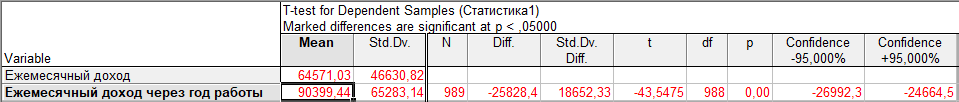
Так лет в компании и всего рабочих лет не являются повторными измерениями одной и той же величины (зависимыми) и измерены в одних и тех же единицах воспользуемся данным методом. Проверим равенство средних для лет в компании и всего рабочих лет на основе t-критерия (t-test, independent, by variables).



Так как уровень значимости р = 0,00 критерия Стьюдента меньше, чем 0,05, то отличие средних переменных статистически значимо, т.е. не верна гипотеза о равенстве средних генеральных совокупностей. И гипотеза о равенстве дисперсии не справедлива, что снижает достоверность вывода.

**3.3.3 t-test, dependent samples**

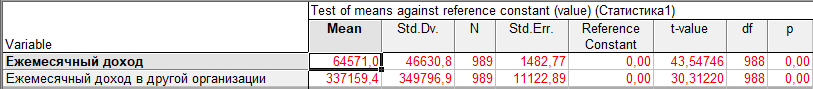
Предположим, что ежемесячный доход и ежемесячный доход через год работы зависимы. Тогда проверим равенство средних для этих переменных на основе t-критерия (t-test, dependent samples).



По данным таблицы можно сделать вывод, что средние отличаются существенно и гипотеза о равенстве дисперсий не справедлива.

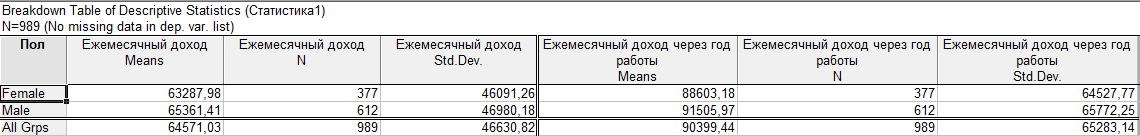
**3.3.4 t-test, single samples**

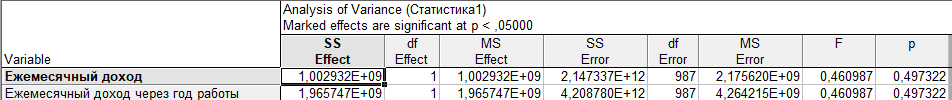
Проверим равенство средних для ежемесячного дохода сотрудника в организации и ежемесячного дохода сотрудника в других организациях на основе t-критерия (t-test, single samples). Этот метод не учитывает дисперсию и может получиться более верный результат.



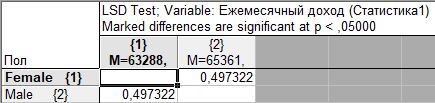
По данным таблицы можно сделать вывод, что средние отличаются существенно.

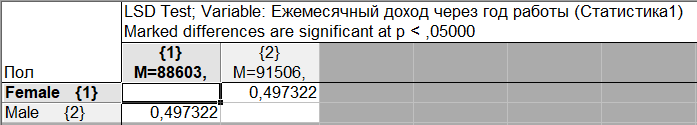
**3.4 Группировка и однофакторная ANOVA**



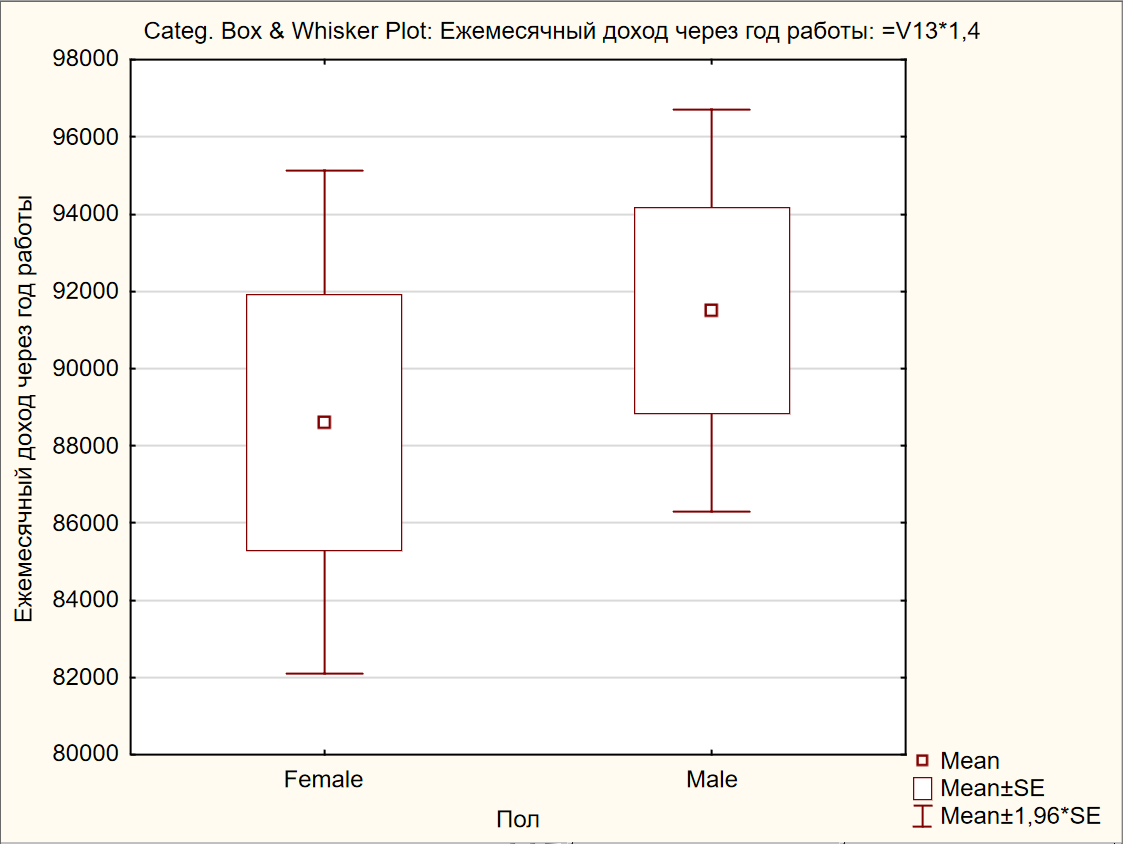


Из таблицы с результатами дисперсионного анализа следует, что для данных переменных отсутствуют группы со статистически значимым отличием. Дополнительно проверим данное утверждение при помощи критерия LSD (НЗР).



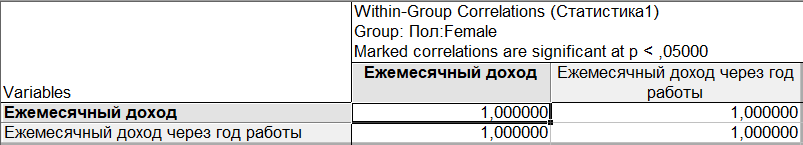


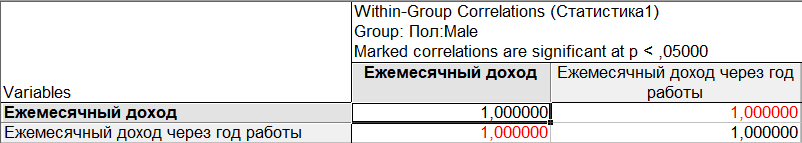




Из таблиц с уровнями значимости р следует подтверждение ранее сделанных выводов об отсутствии групп со статистически значимым отличием по переменным ежемесячный доход и ежемесячный доход через год работы.

Внутригрупповые корреляции:





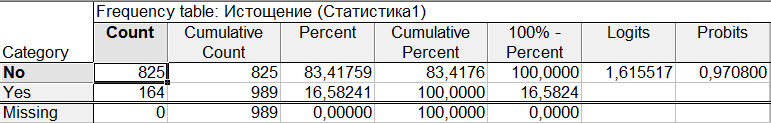
Из таблиц видно, что в разных группах зависимости между анализируемыми величинами проявляются по-разному. Во втором случае зависимость статистически значимая (р<0,05), так как выделены красным цветом. В обоих случаях взаимосвязь сильная, так как показатель выше 0,75 и корреляции положительные, то есть при увеличении одной из переменных, вторая также будет расти.

Написать какие взаимосвязи умеренные, сильные, что означает положительная и отрицательная корреляция!

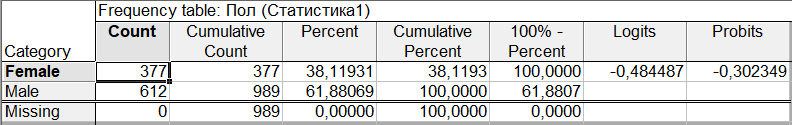
**4 Частотный анализ**

**4.1 Таблицы частот**

Таблицы частот позволяют анализировать категориальные переменные.

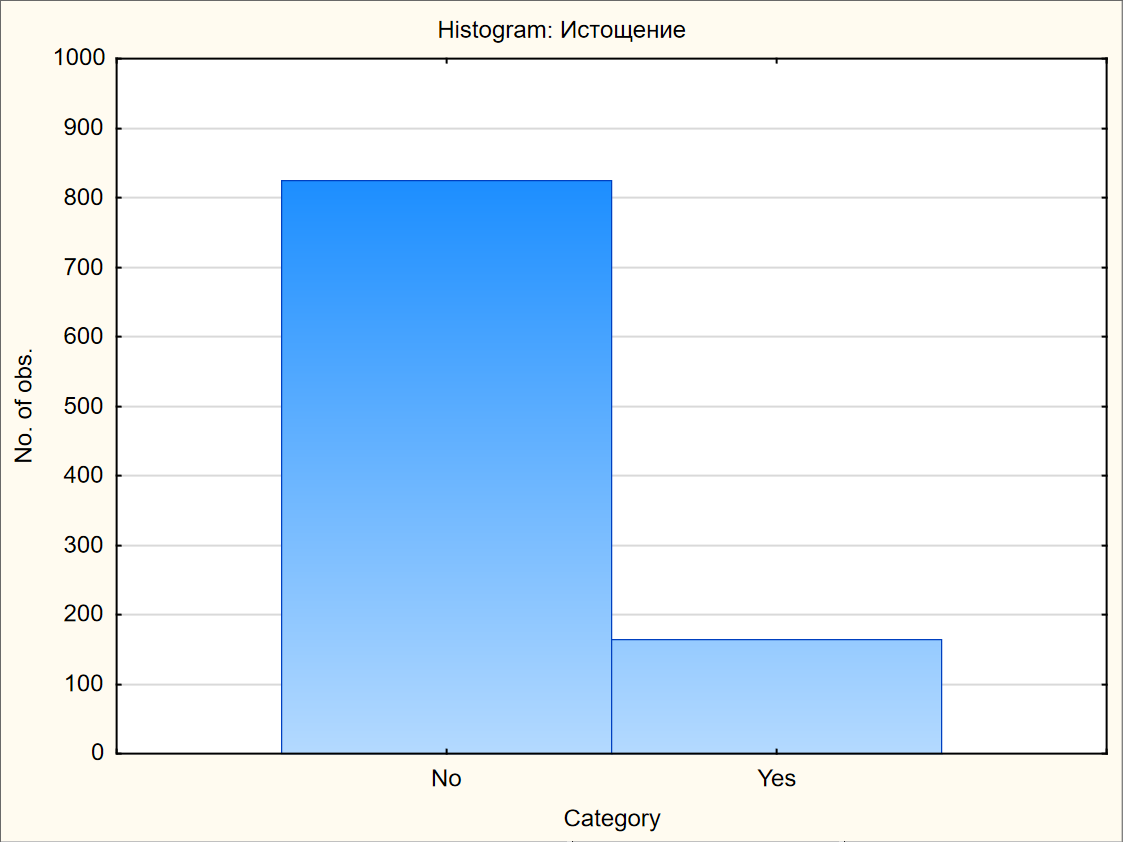


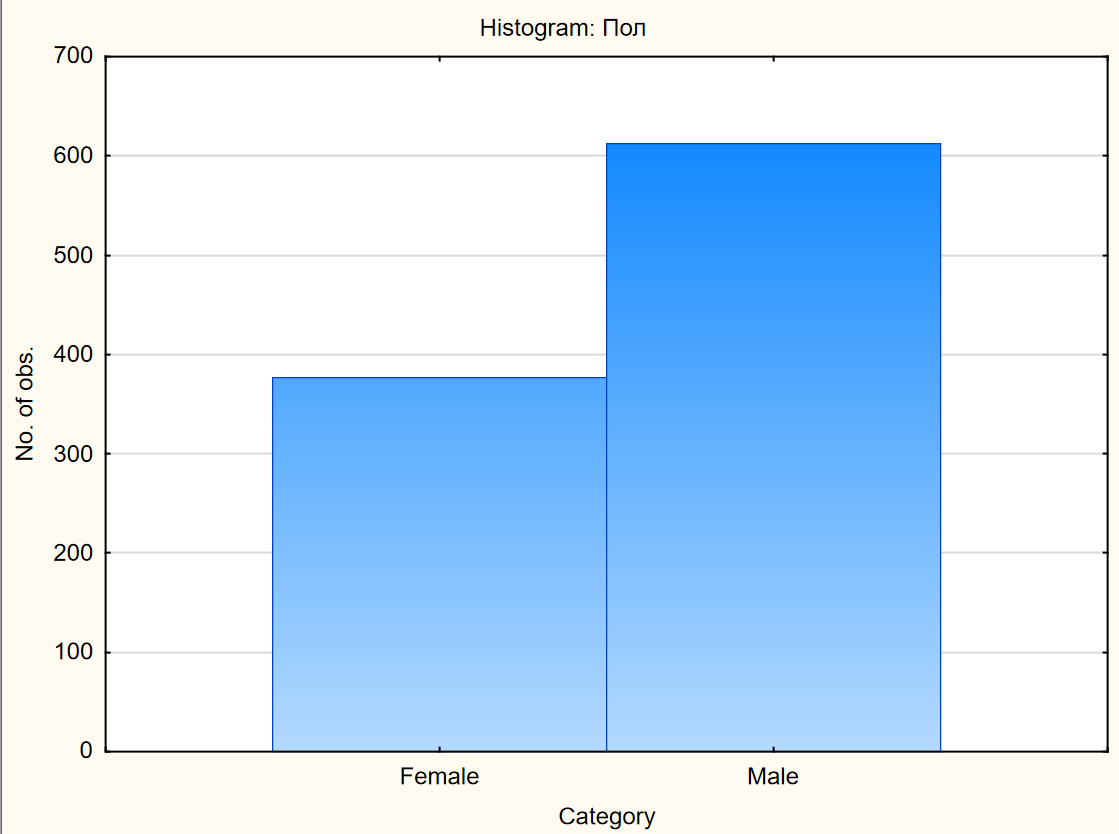
Данная таблица частот построена для категориальной переменной Истощение. В таблицы для категорий No и Yes указаны слева направо: частоты, кумулятивные частоты, проценты (относительные частоты), кумулятивные проценты, 100%-кумулятивные проценты, логит преобразование частот, пробит преобразование частот.



Данная таблица частот построена для категориальной переменной Пол. В таблицы для категорий Female и Male указаны те же показатели, что и в предыдущей таблице.

Гистограммы соответствующих категориальных переменных:





Из гистограмм видно, что больше сотрудников мужского пола и большинство сотрудников не истощены работой.



Данная таблица считает частоты для заданных пользователем категорий. Каждая из этих категорий определяются интервалами: v6>=0 and v6<=25, v6>25 and v6<=50, v6>50 and v6<=100, где v6 – это переменная юрлиц.

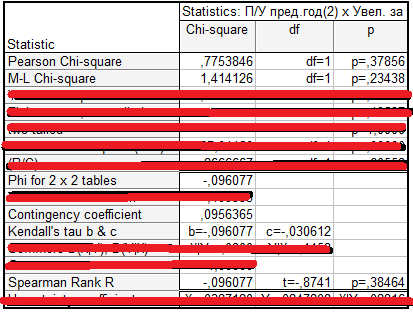
**4.2 Таблицы кросстабуляции**

Таблицы кросстабуляции объединяют несколько таблиц частот.



В данной таблице представлены частоты уже четырех групп, образовавшихся за счет комбинаций всех категорий. Анализ процентов отдельно по строкам или столбцам таблицы показывает, что тренд изменения процентов в разных группах примерно одинаков для разных строк (столбцов), это говорит о том, что нет взаимосвязи или слабая взаимосвязь между строками и столбцами таблицы, а значит и между категориальными переменными. Для оценки статистической значимости и боле достоверного вывода о силе взаимосвязи воспользуемся статистиками, приведенными в таблице

Таблица статистик для двухвходовой таблицы:



Критерии Хи-квадрат Пирсона и наибольшего правдоподобия (M-L) Хи-квадрат проверки статистической значимости связи между двумя группирующими (категориальными) переменными, показывают, что взаимосвязи между строками и столбцами таблицы частот, т.е. между категориальными переменными статистически не значимы, так как уровни значимости р больше, чем 0,05. Значения всех статистик меньше, чем 0,25 и близки к 0, поэтому можно говорить о слабой статистически не значимой взаимосвязи, или ее отсутствии, т.е. справедливости нулевой гипотезы

# 5 Корреляционный анализ

Нулевая гипотеза – коэффициент корреляции равен 0.

Коэффициент Спирмена используется, если закон распределения переменных неизвестен или не является нормальным.



Так как p<0.05, то нулевая гипотеза отвергается, то есть корреляция между переменными статистически значима.

Коэффициент Кендалла используется, если хотя бы одна переменная качественная.



Так как p>0.05, то нулевая гипотеза принимается, то есть корреляция между переменными статистически не значима.

Коэффициент Гамма используют, если переменные содержат много повторяющихся значений.



Так как p<0.05, то нулевая гипотеза отвергается, то есть корреляция между переменными статистически значима.

**6 Дисперсионный анализ**

Дисперсионный анализ является наиболее общим методом сравнения средних. В дисперсионном анализе можно исследовать зависимость количественного признака (зависимой переменной) от одного или нескольких качественных признаков (факторов).

В качестве зависимой переменной выбраны итоги по страхованию имущества и выплаты, в качестве категориальных Прибыль/Убыток за предыдущий год и Увеличилась за год.



Из графика видно, что итоги по страхованию имущества больше у прибыльных страховщиков, чем у убыточных; у прибыльных страховщиков с увеличенной за год прибылью выплаты больше итогов; выплаты у прибыльных страховщиков больше убыточных.



Из таблицы видно, что верна гипотеза о равенстве средних во всех группах, то есть средние итогов страхования равны в этих группах.



Из таблицы видно, что по критериям Кохрана, Хартли и Бартлетта гипотеза однородности дисперсии отклоняется для всех переменных.



Из гистограммы видно, что распределение Sum не соответствует нормальному закону.



Из таблицы видно, что по критерию Левена гипотеза однородности дисперсии отклоняется только для переменной Вып.





Из диаграмм рассеяния видно, что средние и стандартные отклонения для переменной Sum коррелируют существенно, для переменной Вып. коррелируют меньше.

**7 Линейное многомерное моделирование взаимосвязей**

Построим зависимость между итогами по страхованию и остальными параметрами страховщиков.



Из таблицы видно, что коэффициента регрессии при переменных С-М риски и Фин риски статистически значимы (p-level<0.1), остальные статистически незначимы. Зависимость между откликом и предикторами не высокая (R2=0.415), **где информация по коэффициенту множественной корреляции?**



Из данной таблицы видно, что в целом уравнение регрессии статистически значимо.



По данной таблице можно ранжировать влияния предикторов на отклик. Больше всего влияет С-М риски. Этот же предиктор имеет наибольшую частную и получастную корреляции. R-square говорит о множественной корреляции между данной переменной и остальными переменными. Видно, что все коэффициенты детерминации умеренные???. Tolerance=1- R-square. p-level говорит о значимости частных коэффициентов корреляции????? переменных С-М риски, Фин риски и Вып при уровне значимости p=0.1.



Из таблицы видно, что статистика Дарбина-Уотсона имеет небольшое значение (1,9) при низкой сериальной корреляции (зависимость остатков) (0,048).

Графическое сравнение предсказанных и наблюдаемых значений отклика приведено ниже:



Пример прогноза отклика при соответствующих значений предикторов:





По данной гистограмме можно судить о соответствии закона распределения остатков нормальному закону.

Можно сделать вывод о невысокой адекватности построенной линейной модели зависимости итогов страхования имущества от остальных количественных предикторов.

**Интерпретация результатов неточная, есть ошибочные выводы, см. лекции. Не выписано уравнение множественной регрессии**

**8 Нелинейное многомерное моделирование взаимосвязей**

**8.1 Fixed Nonlinear Regression**

Этот модуль реализует множественный линейный регрессионный анализ с линеаризованной моделью.

Построим квадратичную зависимость Выплат от количества заключенных договоров и количества выплат.

Предварительно, до включения в модель множественной регрессии, можно просмотреть корреляции предикторов и функции отклика:



Из таблицы видно, что Выплаты зависят только от количества выплат.



Как видно, коэффициент детерминации высокий (R2=0.868). В модель не включена переменная кол. дог.

Для проверки адекватности модели построим гистограмму остатков.



Видно, что распределение остатков не соответствуют нормальному.

**Интерпретация результатов неточная, есть ошибочные выводы, см. лекции. Не выписано уравнение множественной регрессии**

**8.2 Логит регрессия**

Построим Логит модель зависимости бинарной переменной Увеличилась за год от Доли юрлиц и Изменения за год.



Регрессия значима, так как p<0.05 (отклоняется гипотеза об отсутствии связи).

Нормальный вероятностный график остатков. Чем точки ближе к прямой, тем лучше.



Наблюдаемые и предсказанные:



Видно, что модель правильно предсказывает только страховщиков с увеличенной за год прибылью.

**Интерпретация результатов неточная, есть ошибочные выводы, см. лекции. Не выписано уравнение логит регрессии**

**~~8.3 Пробит регрессия~~**

Построим Пробит модель зависимости бинарной переменной Прибыль/Убыток за предыдущий год от Доли юрлиц и Изменения за год.



Регрессия незначима, так как p>0.05 (отклоняется гипотеза об отсутствии связи).

Нормальный вероятностный график остатков. Чем точки ближе к прямой, тем лучше.



Наблюдаемые и предсказанные:



Видно, что модель правильно предсказывает только прибыльных страховщиков.

**8.4 Экспоненциальная регрессия**

Построим экспоненциальную модель зависимости Выплат от Количества выплат:



Таблица предсказанных и наблюдаемых значений:



Гистограмма остаткова:



Видно, что распределение остатков не соответствует нормальному.

**Интерпретация результатов неточная, есть ошибочные выводы, см. лекции. Не выписано уравнение регрессии**

**~~8.5 Кусочно-линейная регрессия~~**

Построим кусочно-линейную модель зависимости Выплат от Количества выплат:



Из таблицы видно, что доля объясненной дисперсии большая (89,9%), точка разрыва, определенная программой самостоятельно 5028,036.



Из гистограммы видно, что распределение остатков не сильно соответствует нормальному распределению.

Таблица предсказанных и наблюдаемых значений:



**8.6 Определенная пользователем регрессия**

Построим следующую модель: Выплаты=b0+b1\*sqrt(Кол. выплат)



Из таблицы видно, что оба параметра значимы.

Таким образом регрессия имеет вид:

Выплаты=8610,889-38,1\*sqrt(Кол. выплат)

Нормальный вероятностный график остатков. Чем точки ближе к прямой, тем лучше?????.





Из данной таблицы видно, что в целом регрессия значима????.

**Интерпретация результатов неточная, есть ошибочные выводы, см. лекции.**

**~~9 Канонический анализ~~**

Канонический анализ позволяет исследовать зависимость между двумя множествами переменных. Проведем анализ для следующих списков переменных: 1) Итоги по страхованию (Sum), Количество заключённых договоров (Кол. дог.) и 2) Выплаты (Вып), Количество выплат (Кол. вып.).



По данной таблице видно, Каноническая корреляция R=0.84, ее значение свидетельствует о наличии сильной зависимости между группами переменных. Chi-square=100.71 и p=0.00 показывают значимость R. Доля дисперсии, объясняемая каждым множеством равна 100%, реальная изменчивость в левом множестве объясняется 47,63% правого множества, а реальная изменчивость в правом множестве объясняется 35,99% левого множества.

Собственные значения канонических корней представлены в следующей таблице:



Каждая пара канонических переменных объясняет уникальную долю изменчивости в этих двух наборах переменных.



На данном рисунке продемонстрирован график убывающих собственных значений.



Из данных, приведенных в таблице, можно сделать вывод о статистической значимости только первого канонического корня. Значит целесообразно рассматривать лишь первую пару канонических переменных.

Корреляции между переменными из одного и из разных множеств:







Внутри множеств зависимость слабая, а наибольшая зависимость между Sum и Кол. вып.

Факторная структура:





В этих таблицах приведены нагрузки канонических факторов, которые показывают корреляции между переменными из множества и соответствующими каноническими переменными.

Извлеченная дисперсия:





Извлеченную дисперсию можно интерпретировать как среднюю долю дисперсии, объясняемой соответствующим корнем. Коэффициенты избыточности (Reddncy) можно интерпретировать как среднюю долю дисперсии, объясняемую в переменных соответствующего множества, исходя из значения корня, при заданных значениях переменных другого множества.

График канонических корреляций (канонические корреляции равны квадратным корням из собственных значений):



Диаграмма рассеяния канонических корреляций:



Это диаграмма рассеяния значений канонических переменных левого и правого множеств, соответствующих первому каноническому корню. Видно, что линейная зависимость плохо просматривается, как и для второго:



Канонические веса:





Из таблиц видно, что для левого множества наибольший вклад в значение первой канонической переменной вносит переменная Кол. дог., а для правого множества – Кол. Вып.

Таким образом, между списками переменных Итоги по страхованию (Sum), Количество заключённых договоров (Кол. дог.) и Выплаты (Вып), Количество выплат (Кол. вып.) есть сильная зависимость, и наибольший вклад в зависимость вносят переменные Кол. дог. и Кол. Вып.

**10 Дискриминантный анализ**

**Переделать все, анализ проведен неверно, применен другой метод!!!! Надо использовать кластеризацию методом к-средних!**

Цель дискриминантного анализа – на основе измерения различных характеристик объекта классифицировать его, то есть отнести к одной из нескольких групп некоторым оптимальным способом.

Задача состоит в том, чтобы по результатам рассмотрения всех показателей страховщиков отнести их либо к прибыльным, либо к убыточным.



Значение лямбды Уилкса, лежащее около 0, говорит о хорошей дискриминации, около 1 - о плохой. Чем ближе значение показателя предиктора к 1, тем более значим предиктор в модели. Например, в 1-й модели наиболее значим предиктор Увел. за год(0.918).



Из таблицы видно, что канонический корень статистически незначим.



В данной таблице показаны коэффициенты исходной дискриминантной функции и соответствующее ей собственное значение.



В данной таблице показано количество ошибок классификации.

Значимость эффектов предикторов можно увидеть из таблицы:



Как видно из данной таблицы, по критериям Уилкса и Пиллая значимыми будут эффекты предикторов Sum и кол. вып. Это значит, что эти предикторы имеют наибольшее влияние на определение принадлежности наблюдений классам – страховщиков к убыточным или прибыльным.

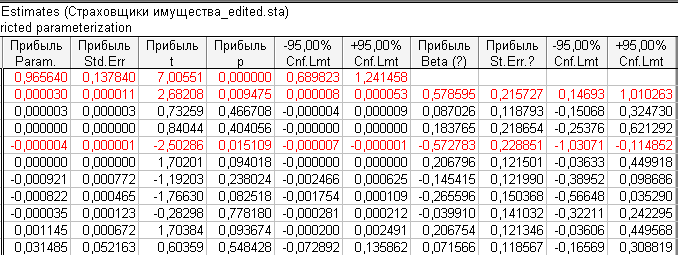
Проект плана:



В этой таблице показаны метки для каждого столбца в матрице планов, названия переменных, уровень и другой уровень.



*продолжение*



В данной таблице приведены оценки параметров, их стандартные ошибки, уровни значимости и соответствующие статистики. Из таблицы видно, что Sum и кол. вып. являются определяющими признаками для убыточных и прибыльных страховщиков. Причем для убыточных параметр для Sum отрицательный, а для кол. вып. положительный, я для прибыльных наоборот (что соответствует здравому смыслу!).

По следующей таблице стандартного дисперсионного анализа можно судить о роли переменных в определении принадлежности наблюдений к тому или иному классу:





В таблице приведены коэффициенты и свободные члены при переменных линейных функций.

Следующая таблица содержит классификации для каждого наблюдения (часть таблицы). Наблюдения, которые не удалось классифицировать помечены знаком «\*»:



В следующей таблице представлены квадраты расстояний Махаланобиса каждого наблюдения от центроида группы, расстояния учитывают корреляции в модели.



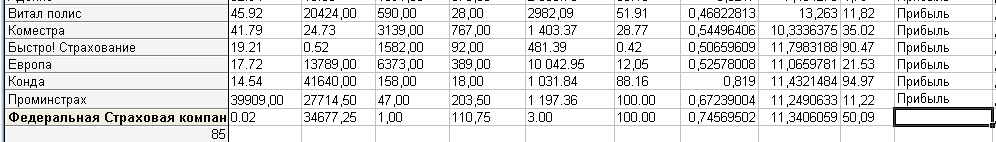


В данной таблице каждому наблюдению поставлена в соответствие вероятность принадлежности к группе.



Из гистограммы видно, что закон распределения остатков напоминает нормальное распределение, поэтому можно сделать вывод о достаточном уровне предсказания для Убыточных страховщиков.

Проведем классификацию для класса, который неизвестен. Для этого удалим в строке 84 значение для столбца П/У пред.год (было Прибыль):



Теперь, проведем еще раз анализ и посмотрим апостериорные вероятности для наблюдения под номером 84:



Из таблицы видно, что наибольшая вероятность, равная 0,966 соответствует прибыльным страховщикам, т.е. программа верно определила сорт цветка.

**11 Классификационный анализ без обучения**

**11.1 Кластерный анализ**

**11.1.1 Метод к-средних**

Кластерный анализ позволяет разбить множество исследуемых объектов и признаков на однородные в некотором смысле группы, или кластеры.

Попробуем разбить страховщиков на кластера, но до этого надо стандартизировать переменные. Варьируя количеством кластеров и исключая переменные, получили следующие результаты:



По графику средних значений для каждого кластера видно, что разбиение на 3 кластера не качественное. Разобьем объекты на 2 кластера:



Из графика видно, что в кластерах средние значения параметров Вып. и Юрлиц незначительно отличаются друг от друга. Попробуем их исключить:





Таблица и график свидетельствуют об успешной классификации страховщиков без учета параметров Вып. и Юрлиц. В таблицы приведены значения межгрупповых и внутригрупповых дисперсий признаков. Чем меньшее значение внутригрупповой и больше межгрупповой дисперсии, тем лучше признак характеризует принадлежность объектов к кластеру. Так как p<0.05 для всех признаков, то вклады всех признаков в разделение объектов на группы существенные???. Вывод неверный, см. лекции





В первой из данных таблиц приведены средние для каждого кластера, во второй – евклидовы расстояния и квадраты евклидовых расстояний между кластерами.

**~~11.1.2 Двухвходовая кластеризация~~**

Проведем двухвходовую кластеризацию (одновременно и по переменным и по строкам) для всех количественных признаков всех страховщиков.

Ниже приведен цветной график результата кластеризации:



Из графика видно, что, например, 18-й страховщик и 77-й попали в один кластер по Sum, 57-й и 68-й по Вып. и т.д.



В данной таблице приведены значения средних и стандартных отклонений для строк. (Для переменных аналогичные показатели равны 0 и 1 соответственно, т.к. переменные стандартизированы).

**11.1.3 Древовидная кластеризация**

Проведем древовиднуюкластеризацию по переменным.

Горизонтальная древовидная диаграмма имеет вид:



На диаграмме видно, что по мере понижения порога (изменение расстояния объединения), относящегося к решению об объединении объектов в один кластер, все большее и большее число объектов связывается, и объединяются все больше кластеров.



На данном графике показан порядок объединения. Например, на 2-м шаге расстояние объединения должно было значительно вырасти, чтобы объединить еще переменные.

Правило объединения в кластеры показано в следующей таюлице:



**Кластеризация проведена неверно, надо кластеризовать по объектам, а не переменным! Переделать!**

**11.2 Деревья классификации**

Деревья классификации – это метод, позволяющий предсказывать принадлежность наблюдений или объектов к тому или иному классу категориальной зависимой переменной в зависимости от соответствующих значений одной или нескольких независимых переменных.

Будим классифицировать страховщиков на прибыльных и убыточных в зависимости от выплат.



Из таблицы видно, что левая и правая ветви содержат по одному узлу с номерами 2 и 3 соответственно. Из строки 1 видно, что в первой вершине все страховщики классифицированы как Убыточные. Условие разделения страховщиков по вершинам 2 и 3 следующее: если значение кол. вып.<=4061.50, то страховщики классифицируются как Прибыльные, остальные как убыточные.

Граф дерева классификации изображен на следующем рисунке:



Условия ветвления:



Последовательность деревьев:



В таблице указаны терминальные вершины, цена кросс-проверки, ее стандартная ошибка, цена обучения и сложность каждого из усеченных деревьев. Звездочкой помечено дерево, которое было признано деревом «подходящего размера».

График последовательности цен:





В данной таблице показана значимость предиктора.



В данной таблице показана предсказанные и наблюдаемые объекты в классах.

На следующем рисунке показана дискретная карта линий уровня:



Априорные вероятности:



В этой таблице показаны априорные вероятности для каждого класса зависимой переменной и число элементов в каждом классе.

Ошибки классификации:



В этой таблице показано, сколько объектов каждого класса было ошибочно отнесено к другому классу.

В следующей таблице представлена матрица ошибок классификации глобальной кросс-проверки:



В данной таблице видно, что 35 убыточных страховщиков неверно классифицированы как прибыльные, 3 прибыльных страховщиков неверно классифицированы как убыточные. Кроме того, в таблице показана цена глобальной кросс-проверки и ее стандартное отклонение: 0,47436 и 0,10588 соответственно.

Таким образом, получено решающее правило, состоящее из одного этапа, которое позволит произвольного страховщика классифицировать как прибыльного или убыточного по значению количества выплат.

**Дерево классификации построено неправильно, нет верных классификаций, переделать!!!**

**12 Методы редукции данных**

**12.1 Факторный анализ**

Главная цель факторного анализа это сокращение числа переменных и определение структуры взаимосвязей между переменными, т.е. классификация переменных.



Из таблицы видно что Sum больше коррелирует с кол. дог., кол. вып. и фин. риски. Анализ проведем по переменным: Sum, Вып., кол. дог., кол. вып. и юрлиц.

Просмотрим собственные значения факторов:



Во втором столбце (Eigenvalue) приведены дисперсии выделенных факторов – собственные числа, в третьем – процент от общей дисперсии. Как видно, первый фактор объясняет 45% общей дисперсии, второй – 26% и т.д. Необходимо понять, сколько факторов следует оставить.

По критерию Кайзера можем отобрать только факторы с собственными значениями, большими 1. Из таблицы видно, что на основе данного критерия выделяются только 2 фактора.

Критерий каменистой осыпи является графическим:



Надо выбрать такое место на графике, где убывание собственных значений слева направо максимально замедляется. Из графика видно, что в соответствии с этим критерием можно попытаться выделить 2 или 3 фактора.

Предположим, что число факторов неизвестно.

Корреляции между переменными и выделенными факторами:



Из таблицы видно, что первому и второму факторам соответствуют большие коэффициенты корреляции, чем остальным. Назначим число факторов 2.



Из таблицы видно, что Factor 1 имеет высокие факторные нагрузки по переменным Sum, Кол. дог. и Кол. вып. и низкие по переменным Вып. и Юрлиц. А Factor 2 наоборот. Это означает, что выделенные два фактора наилучшим образом характеризуют данные.



Этот график иллюстрирует соотношение между факторами и группами переменных. Видно, что группа переменных Вып. и Юрлиц занимает на плоскости крайнее левое верхнее положение, группа переменных Sum, Кол. дог. и Кол. вып. крайнее правое нижнее положение.

Коэффициенты уравнений регрессий, по которым программа посчитает значения факторов для каждого наблюдения, представлены в таблице:



Значения факторов для каждого страховщика представлены в таблице:



По данным из этой таблицы можно судить об отношении страховщиков к соответствующему фактору. Положительное значение фактора соответствует позитивному отношению страховщика, а отрицательное – негативному.

**~~12.2 Метод анализ главных компонент и классификация~~**

Метод анализ главных компонент и классификация служит для достижения двух целей: уменьшение общего числа переменных и классификация переменных и наблюдений, при помощи строящегося факторного пространства.

Выберем переменные:

для анализа - Sum, Вып., Кол дог., Кол. вып.;

вспомогательные переменные – Страх. Сум., Юрлиц, Фин риски, С-М риски, Изм.;

переменные с основными наблюдениями – Увел. за год.;

группирующая переменная – П/У пред. год.

График каменистой осыпи:



Из графика видно, что число выделяемых факторов может быть 2 или 3.



Эта таблица собственных значений. Во втором столбце (Eigenvalue) приведены дисперсии выделенных факторов – собственные числа, в третьем – процент от общей дисперсии. Как видно, первый фактор объясняет 55,6% общей дисперсии, второй – 26% и т.д. По критерию Кайзера можем отобрать только факторы с собственными значениями, большими 1. Из таблицы видно, что на основе данного критерия выделяются только 2 фактора.



В данной таблице представлены факторные координаты переменных (факторные нагрузки). Большее абсолютное значение факторной нагрузки переменной с каким-либо фактором говорит о том, что переменная сильнее связана с этим фактором. Вспомогательные переменные обозначены «\*». Первая факторная ось, соответствующая собственному значению 2,22, наиболее сильно коррелирует с переменными Кол. дог. и Кол. вып. (сильные положительные корреляции), Sum (умеренные положительные корреляции). Вторая факторная ось, соответствующая собственному значению 1,04, наиболее сильно коррелирует с переменной Вып. (сильные отрицательные корреляции).

График факторных координат переменных и наблюдений:



Чем ближе переменная к единичной окружности, тем лучше она воспроизведена в найденной системе координат (лучше воспроизводится текущим набором выделенных факторов).

Вклад переменных в дисперсию факторной оси:





Это таблица общностей переменных. Общность – это доля объясненной дисперсии, которая характеризует степень общности переменной с другими переменными по заданному числу факторов. Из таблицы общностей переменных видно, что самая высокая степень общности с другими переменными для первого фактора у переменной Кол. вып., у второго фактора - у Вып.



В данной таблице представлена информация о принадлежности наблюдения к основным или вспомогательным наблюдениям. Например, страховщики Согласие и Национальная страховая группа относятся к не основным наблюдениям.

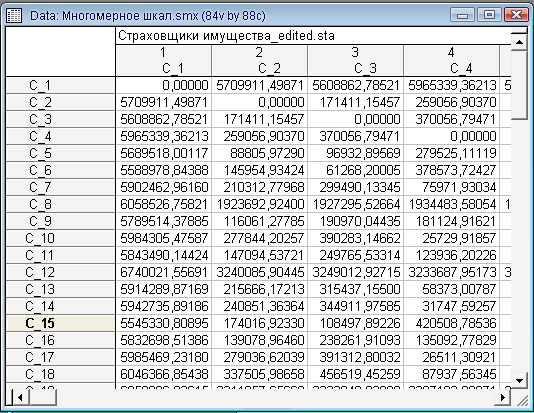
График наблюдений в факторном пространстве:

 Наблюдения синего цвета - основные (страховщики, у которых прибыль увеличилась за год), красного - вспомогательные (страховщики, у которых прибыль не увеличилась за год).

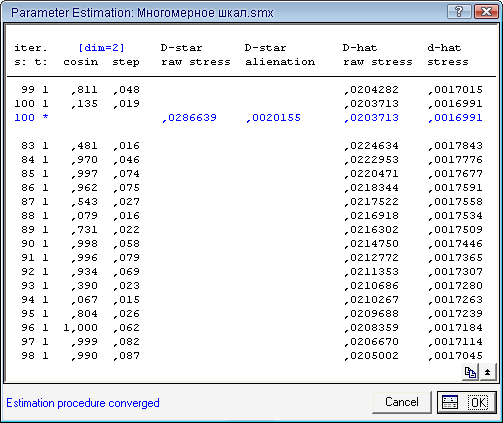
**~~13 Многомерное шкалирование~~**

Цель многомерного шкалирования – поиск и интерпретация латентных переменных, дающих возможность пользователю объяснить сходства между объектами, заданными точками в исходном пространстве признаков.

Перед работой надо создать матричный файл. Его фрагмент представлен ниже:



В результате оценивания параметров получим:



Из этой таблицы видно, что на 100-ой итерации получили лучшую двухмерную конфигурацию. D-star вычисляется с помощью процедуры, которая пытается воспроизвести порядок следования рангов расстояний в исходной матрице расстояний. D-hat вычисляется с помощью метода, который позволяет подобрать монотонное преобразование (регрессию), наиболее точно воспроизводящее исходные расстояния.



В этой таблице содержаться воспроизведенные расстояния и соответствующие им значения D-hat и D-star. Для обозначения элементов используют матричный индекс, например, D(84,1) – 84 это строка матрицы входов, 1-колонка, то есть определяется расстояние между Федеральной Страховой компанией и СГ Росгосстрах.

Таблица результатов с координатами окончательной конфигурации:



Таблица результатов с преобразованными входными значениями, вычисленными с помощью монотонной регрессии:

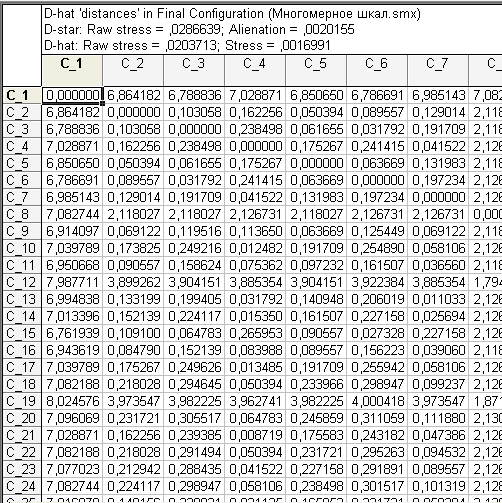
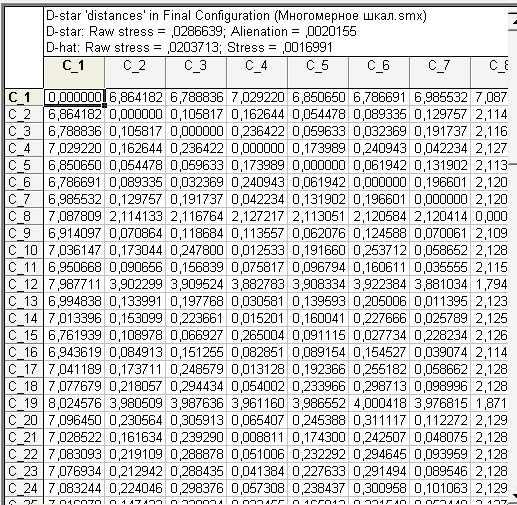
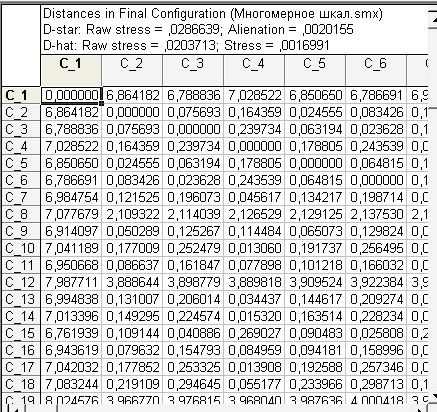


Таблица результатов с преобразованными входными значениями, вычисленными с помощью процедуры ранговых образов Гутмана.



Матрица расстояний для конфигурации точек в пространстве размерности 2.





На этой диаграмме изображена зависимость воспроизведенных расстояний от исходных расстояний. Чем лучше согласие ступенчатой функции с точками данных на диаграмме рассеяния, тем лучше согласие с моделью (воспроизведение расстояний).



Это график окончательной конфигурации объектов на плоскости. Из графика видно, что страховщики 1 и 8 «далеки» от остальных по определенным характеристикам.

График зависимости преобразованных значений входных данных от преобразованных расстояний:



Чем плотнее точки вокруг диагональной линии, тем лучше согласие с данными для выбранной модели.

График зависимости преобразованных значений входных данных от преобразованных расстояний:

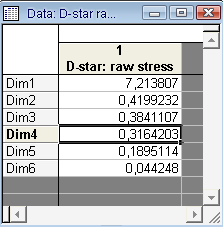


График окончательной конфигурации объектов в трехмерном пространстве.

Если выбрать размерность 3, то можем просмотреть график окончательной конфигурации объектов в трехмерном пространстве:



Для проверки правильности выбора размерности пространства воспользуемся критерием каменистой осыпи. Для этого проведем последовательно вычисления D-star: raw stress для размерностей от 6 до 1, запишем полученные данные в таблицу и построим линейный график по полученной таблице.





Из графика согласно критерию каменистой осыпи следует, что для воспроизведения расстояний между страховщиками необходимо выбрать двухмерное пространство, так как в точке с абсциссой Dim2 максимально замедляется уменьшение стресса.