Лабораторная работа №1.

1.1. Описательная статистика. Корреляционный анализ

Цель работы: изучение корреляционного анализа и явлений ложной корреляции.

Среда выполнения: MS Excel, LibreOffice Calc, Statistica, RStudio, SPSS, Deductor (по выбору студента).

Задание

1. Выбрать массив данных (ссылки представлены ниже в списке литературы, рекомендуется №7), описать параметры, зависимые и независимые переменные. Примечание: при переносе данных на лист MS Excel необходимо использовать импорт данных (учитывая разделитель столбцов и тип столбцов).
2. Провести дескриптивный анализ, оценить близость выборок к нормальной.
3. Построить гистограммы выбранных атрибутов. Длину интервала рассчитать с помощью формулы Стерджесса.
4. Изучить пример построения таблиц сопряженности (п. 4 из списка литературы). Выбрать в рассматриваемой выборке категориальные переменные или перевести числовые переменные в категориальные, провести анализ с помощью таблиц сопряженности. Описать результаты.
5. Провести корреляционный анализ. Если рассматриваемые выборки далеки от нормального распределения – провести ранговый корреляционный анализ.
6. Оценить значимость корреляции.
7. Выделить сильно- и слабокоррелированные признаки.
8. Интерпретировать результаты.
9. На сайте http://www.tylervigen.com/spurious-correlations выбрать массив данных, построить графики, оценить ложную корреляцию между параметрами.
10. Оформить отчет.

Содержание отчета

1. Титульный лист.
2. Цель работы.
3. Описание исходных данных.
4. Результаты дескриптивного анализа.
5. Анализ результатов таблиц сопряженности.

6. Результаты корреляционного анализа.
7. Оценка и интерпретация результатов.
8. Пример и интерпретация ложной корреляции.

Список литературы и ссылки на материалы

1. Айвазян С.А. Методы эконометрики. – М.: Магистр: - ИНФРА-М, 2010. – 512 с.
2. Группировка и формула Стерджесса. http://univer-nn.ru/zadachi-po-statistike-primeri/gruppirovka-formula-sterdzhessa/
3. Проверка статистических гипотез в MS Excel. http://excel2.ru/articles/proverka-statisticheskih-gipotez-v-ms-excel
4. Таблицы сопряженности. http://excel2.ru/articles/kriteriy-nezavisimosti-hi-kvadrat-v-ms-excel
5. http://statsoft.ru/home/textbook/glossary/GlossaryTwo/P/PearsonCorrelation.htm?sphrase_id=67490
6. https://www.kdnuggets.com/datasets/index.html
7. http://archive.ics.uci.edu/ml/datasets.html
8. http://www.tylervigen.com/spurious-correlations

Вопросы к защите

1. Коэффициент корреляции, его свойства. Корреляция ранговых переменных. Корреляция числовых переменных.
2. Независимость и некоррелированность переменных.
3. Оценка близости выборки к нормальной.
4. Гистограмма.
5. Таблицы сопряженности.
6. Ложная корреляция.

1.2. Регрессионный анализ

Цель работы: изучение регрессионного анализа.

Среда выполнения: MS Excel, LibreOffice Calc, Statistica, R, SPSS, Deductor (по выбору студента).

Задание

1. Выбрать массив данных (ссылки представлены ниже), описать параметры, зависимые и независимые переменные. Можно использовать массив из первой части лабораторной работы №1. Примечание: при переносе данных на лист MS Excel необходимо использовать импорт данных (учитывая разделитель столбцов и тип столбцов).

Пункты 2 - 7 (и соответствующие им пункты отчета) выполняются, если выбран новый массив данных.

2. Провести дескриптивный анализ, оценить близость выборок к нормальной.
3. Построить гистограммы выбранных атрибутов. Длину интервала рассчитать с помощью формулы Стерджесса.
4. Провести корреляционный анализ. Если рассматриваемые выборки далеки от нормального распределения – провести ранговый корреляционный анализ.
5. Оценить значимость корреляции.
6. Выделить сильно- и слабокоррелированные признаки. Построить диаграммы рассеивания для выбранных признаков.
7. Интерпретировать результаты.
8. Построить линейную регрессионную модель для признаков с высоким коэффициентом корреляции.
9. Вывести график остатков. Оценить постоянство среднего и дисперсии.
10. Построить гистограмму стандартизированных остатков.
11. Записать уравнение регрессии.
12. По значениям коэффициента детерминации, графику и гистограмме остатков оценить качество построенной модели.
13. Оценить значимость построенной модели.
14. Дополнить отчет.

Содержание отчета

1. Титульный лист.
2. Цель работы.
3. Описание исходных данных.
4. Результаты дескриптивного анализа.
5. Результаты корреляционного анализа. Диаграммы рассеивания.
6. Регрессионная модель. Уравнение регрессии, график остатков, гистограмма остатков. Оценка качества построенной модели.
7. Интерпретация результатов.

Список литературы и ссылки на материалы

1. Айвазян С.А. Методы эконометрики. – М.: Магистр: - ИНФРА-М, 2010. – 512 с.
2. Группировка и формула Стерджесса. http://univer-nn.ru/zadachi-po-statistike-primeri/gruppirovka-formula-sterdzhessa/
3. Проверка статистических гипотез в MS Excel. http://excel2.ru/articles/proverka-statisticheskih-gipotez-v-ms-excel
4. http://statsoft.ru/home/textbook/glossary/GlossaryTwo/P/PearsonCorrelation.htm?sphrase_id=67490

5. Регрессия в MS Excel. http://exceltip.ru/%D1%80%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D1%8F-%D0%B2-excel/
6. Регрессия в MS Excel. http://archie-goodwin.net/load/specializirovannye_blogi/ms_office/linejnaja_regressija_v_excel_cherez_analiz_dannykh/28-1-0-391
7. https://www.kdnuggets.com/datasets/index.html
8. http://archive.ics.uci.edu/ml/datasets.html

Вопросы к защите

1. Коэффициент детерминации.
2. Уравнение регрессии.
3. Диаграмма рассеяния.
4. Оценка качества регрессионной модели.
5. Виды регрессионных моделей.

1.3. Дисперсионный анализ

Цель работы: изучение однофакторного дисперсионного анализа.

Среда выполнения: MS Excel, LibreOffice Calc, Statistica, R, SPSS, Deductor (по выбору студента).

Задание

1. Выбрать массив данных (ссылки представлены ниже), описать параметры, зависимые и независимые переменные. Можно использовать массивы из лабораторных работ №1 и №2. Примечание: при переносе данных на лист MS Excel необходимо использовать импорт данных (учитывая разделитель столбцов и тип столбцов).

   Пункты 2 - 3 (и соответствующие им пункты отчета) выполняются, если выбран новый массив данных.
2. Провести дескриптивный анализ, оценить близость выборок к нормальной.
3. Построить гистограммы выбранных атрибутов. Длину интервала рассчитать с помощью формулы Стерджесса.

4. Сформулировать начальную гипотезу.
5. Для независимой переменной (фактора) определить градации, разбить значения зависимой переменной в соответствии с

градациями фактора. Если фактор – не категориальная переменная, а числовая, аргументировать разбиение ее значений на интервалы.
6. Провести однофакторный дисперсионный анализ.
7. Подтвердить или опровергнуть выдвинутую гипотезу. Интерпретировать результаты.
8. *Дополнительное задание. Провести двухфакторный анализ.*
9. Дополнить отчет.

Содержание отчета

1. Титульный лист.
2. Цель работы.
3. Описание исходных данных.
4. Результаты дескриптивного анализа.
5. Формулировка гипотез.
6. Описание градаций фактора.
7. Результаты дисперсионного анализа.
8. Интерпретация результатов.

Список литературы и ссылки на материалы

1. Айвазян С.А. Методы эконометрики. – М.: Магистр: - ИНФРА-М, 2010. – 512 с.
2. Группировка и формула Стерджесса. http://univer-nn.ru/zadachi-po-statistike-primeri/gruppirovka-formula-sterdzhessa/
3. Проверка статистических гипотез в MS Excel. http://excel2.ru/articles/proverka-statisticheskih-gipotez-v-ms-excel
4. Однофакторный дисперсионный анализ в MS Excel. https://sites.google.com/site/umkmatematosnovyps/home/rodionov-m/matematiko-statisticeskie-metody-resenia-eksperimentalnyh-psihologiceskih-zadac/glava-2-soderzanie-prakticeskih-zanatij-po-kursu-matematiceskie-osnovy-psihologii-/p-15-resenie-prikladnyh-zadac-sredstvami-excel-odnofaktornyj-dispersionnyj-analiz
5. Дисперсионный анализ. http://statsoft.ru/home/textbook/modules/stanman.html
6. Дисперсионный анализ. http://statistica.ru/theory/dispersionnyy-analiz-article/
7. https://www.kdnuggets.com/datasets/index.html
8. http://archive.ics.uci.edu/ml/datasets.html

Вопросы к защите

1. Процедура применения однофакторного дисперсионного анализа.
2. Двухфакторный дисперсионный анализ.
3. Формулировка гипотез.
4. Подтверждение или отклонение гипотез. Критерий Фишера.
5. Область применения дисперсионного анализа.

1.4. Анализ и дополнения

1. **Оценка близости выборки к нормальной**.
   1.1. Проанализировать полученные значения параметров (асимметрия, эксцесс, мода, среднее) и построенные гистограммы. Сформулировать правила (критерии) для определения близости выборки к нормальной, оценки симметричности распределения.
   1.2. Определить, какие из выбранных признаков можно в дальнейшем использовать для построения моделей.
   1.3. Дополнить отчет.

2. **Корреляционный анализ**.
   2.1. Сравнить процедуры проведения корреляционного анализа и полученные коэффициенты для числовых, категориальных и ранговых переменных.
   2.2. Проанализировать соотношение понятий «зависимость» и «корреляция»: в каких случаях некоррелированность эквивалента независимости, в каких случаях независимость эквивалентна некоррелированности.

3. **Сравнение регрессионных моделей**.
   3.1. Построить диаграмму рассеяния для тех же переменных, для которых была построена линейная регрессионная модель в лабораторной работе №2. *Примечание*: зависимая и независимая переменные должны соответствовать выбранным в регрессионной модели! Диаграмма рассеивания должны была быть построена ранее в рамках выполнения заданий лабораторной работы №1 («корреляционный анализ»).
   3.2. Добавить линию тренда на диаграмму рассеяния.
   3.3. Выбрать тренд, который лучше всего подходит для данной зависимости. *Примечание*: в полиномиальной модели можно повышать степень. **Линейную модель выбирать не нужно**.
   3.4. Отобразить уравнение и коэффициент детерминации (не забыть поставить «галочки» в диалоговом окне).
   3.5. Посчитать новые значения зависимой переменной (y) по полученному уравнению.
   3.6. Посчитать остатки в новой модели («y исходное минус y предсказанное (посчитанное по уравнению)»).
   3.7. Построить график остатков.
   3.8. Сравнить две регрессионные модели: линейную из раздела 2 лабораторной работы и построенную в текущем задании (сравнить значения коэффициентов детерминации, графики остатков).
   3.9. Интерпретировать результаты.
   3.10. Добавить результаты в отчет.

4. **Поиск причин ложной корреляции**.
   4.1. Восстановить таблицу данных и построить диаграмму, соответствующие массиву из раздела «ложная корреляция». *Примечание*: вместо отсчетов по годам можно использовать отсчеты «1», «2» и т.д. Если данные двух рядов сильно отличаются друг от друга – использовать вспомогательную ось с другим масштабом (правой

кнопкой мыши щелкнуть на ряд на диаграмме, выбрать «формат ряда данных» и «отобразить по вспомогательной оси»).

4.2. Добавить линии тренда для каждого из рядов. Отобразить уравнения на диаграмме.

4.3. Посчитать новые значения зависимых переменных по построенным моделям.

4.4. Посчитать остатки для каждого из рядов («y исходное минус y предсказанное (посчитанное по уравнению)»).

4.5. Посчитать стандартизированные остатки.

4.6. Построить диаграмму по остаткам, гистограмму по стандартизированным остаткам.

4.7. Найти корреляцию остатков. Можно использовать функцию MS Excel КОРРЕЛ() или «Корреляция» из «Анализа данных».

4.8. Интерпретировать результаты.

4.9. Добавить результаты в отчет по первой лабораторной работе.

5. **Дисперсионный анализ**.

5.1. Сформулировать правила и условия применения дисперсионного анализа.

5.2. Проанализировать результаты регрессионного анализа из ч.2 лабораторной работы (таблица «Дисперсионный анализ»).

5.3. Дополнить отчет.

Перечень вопросов к защите.

1. Оценка близости выборки к нормальной. Среднее, мода, асимметрия, эксцесс. Симметричные и несимметричные распределения. Гистограмма. Связь гистограммы и функции плотности распределения вероятностей.
2. Корреляционный анализ числовых, категориальных и ранговых переменных. Коэффициенты корреляции Пирсона, Спирмена, Кендалла. Причины ложной корреляции.
3. Регрессионные модели, оценка качества построенной модели (по коэффициенту детерминации, графику остатков, гистограмме остатков).
4. Однофакторный дисперсионный анализ. Двухфакторный дисперсионный анализ. Градации фактора. Критерий Фишера. Процедура применения. Условия применения метода.