**Национальный исследовательский университет**

**«Высшая школа экономики»**

Факультет бизнеса и менеджмента

Отчет о самостоятельной работе по

дисциплине «Эконометрика»

**Бригада №\_\_\_:**

ФИО, курс, группа

…

ФИО, курс, группа

Москва 2017

**Оглавление**

[1](#_30j0zll) Общая постановка задачи 3

[1.1](#_1fob9te) Формулировка прикладной проблемы 3

[1.2](#_3znysh7) Потенциальные потребители решения; задачи, которые они смогут решать, используя полученные результаты 3

[1.3](#_2et92p0) Основные гипотезы, которые планируется проверить в рамках решения задачи 3

[1.4](#_tyjcwt) Основные источники данных 3

[1.5](#_3dy6vkm) Публикации по данной тематике (эссе) 3

[2](#_1t3h5sf) Предварительный анализ собранных данных 5

[2.1](#_4d34og8) Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и выбросы 5

[2.1.1](#_2s8eyo1) Анализ количественных переменных 5

[2.2.2.](#_17dp8vu) Анализ качественных переменных. 6

[2.2](#_3rdcrjn) Анализ статистической связи. 6

[2.2.1](#_26in1rg) Графический анализ пары «числовая зависимая переменная – качественная независимая переменная». 6

[2.2.2](#_lnxbz9) Графический анализ пары «числовая зависимая переменная – числовая независимая переменная». 7

[2.2.3](#_35nkun2) Анализ наличия корреляции между независимыми переменными. 7

[2.2.4](#_1ksv4uv) Предварительная проверка гипотез 7

[3](#_44sinio) Спецификация, оценивание и оптимизация модели 8

[3.1](#_2jxsxqh) Связь структуры базовой модели с основными гипотезами исследования. 8

[3.2](#_z337ya) Оценивание базовой модели и результаты проверки гипотез. 8

[3.3](#_3j2qqm3) Анализ наличия выбросов 9

[3.4](#_1y810tw) Анализ наличия гетероскедастичности 9

[3.5](#_4i7ojhp) Оптимизация модели 9

[3.6](#_2xcytpi) Проверка прогностических свойств модели 9

[4](#_1ci93xb) Выводы и рекомендации 9

# Общая постановка задачи

## Формулировка прикладной проблемы

Здесь необходимо сформулировать, какую прикладную задачу вы хотите решить. Например, при заданном диапазоне денежных средств и сформулированных требованиях к характеристикам жилья выбрать вариант квартиры для покупки. Примеры можно посмотреть в образцах работ.

## Потенциальные потребители решения; задачи, которые они смогут решать, используя полученные результаты

Примеры можно посмотреть в образцах работ.

## Основные гипотезы, которые планируется проверить в рамках решения задачи

Необходимо описать характеристики изучаемых объектов и/или явлений:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Характеристика объекта/явления | Название переменной | Шкала измерения | Роль: зависимая/независимая |
|  |  |  |  |  |
|  |  |  |  |  |

В анализе обязательно должно присутствовать не менее двух количественных и двух качественных независимых переменных. Зависимая переменная — количественная.

Здесь необходимо сформулировать гипотезы о статистической взаимосвязи зависимой переменной и независимыми. В число гипотез должна входить хотя бы одна гипотеза о взаимодействии, т.е. предположение о том, что наличие и характер связи одной из независимых переменных и зависимой зависит от другой независимой переменной. Например, «влияние возраста на спрос зависит от пола, мужчины раньше проявляют интерес к данному товару».

## Основные источники данных

Здесь необходимо указать откуда были взяты данные, в частности указать адреса в Интернете.

## Публикации по данной тематике (эссе)

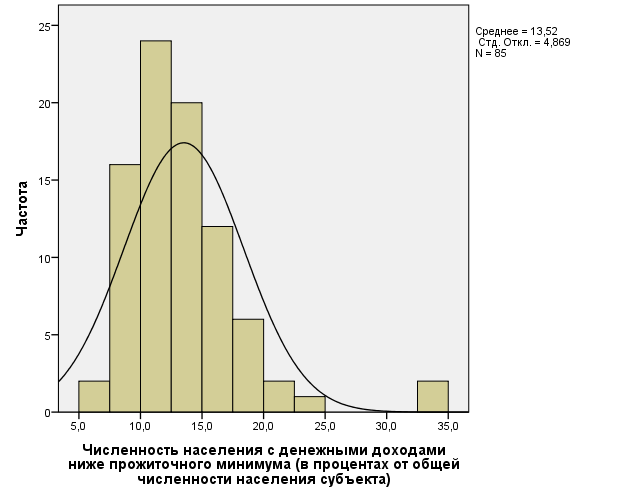
Здесь необходимо указать одну – две публикации на тему исследования с указанием их местонахождения. Тексты публикаций необходимо приложить к отчету. В данном разделе следует также привести краткие (не более одного листа на публикацию) обзоры найденных публикаций. В обзоре следует изложить постановку задачи автора и основные результаты, а также использование публикации в настоящем исследовании.

# Предварительный анализ собранных данных

## Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и выбросы

### Анализ количественных переменных

Здесь необходимо построить и проанализировать гистограммы для всех количественных (интервальных и относительных) переменных в анализе. Необходимо охарактеризовать вид распределения по отношению к нормальному распределению — асимметрию, эксцесс, поли модальность. Для этого следует привести график гистограммы совместно с графиком плотности нормального распределения, а также таблицу основных статистик.

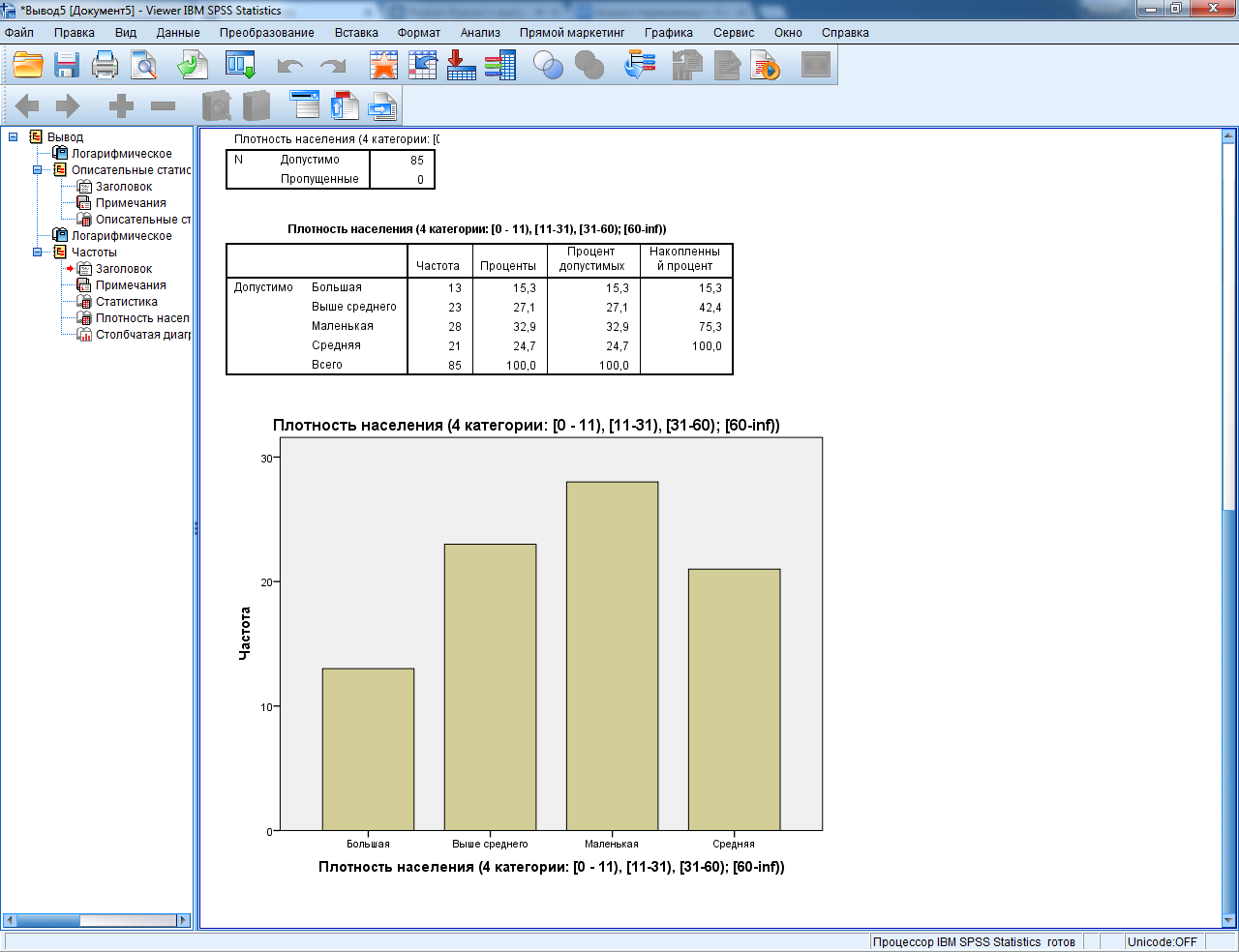


|  |  |
| --- | --- |
| Статистика | Значение |
| Среднее |  |
| Медиана |  |
| Стандартное отклонение |  |
| Межквартильный размах |  |
| Верхняя квартиль |  |
| Нижняя квартиль |  |
| Коэффициент асимметрии |  |
| Коэффициент эксцесса |  |
| Количество наблюдений |  |
| Количество пропущенных значений |  |

Также, здесь следует проанализировать наличие выбросов на основании правила «трех-сигм». Для зависимой переменной следует отметить в базе все выбросы и на основании сравнения значений зависимых переменных с их средними/медианными значениями объяснить, почему эти наблюдения могут интерпретироваться как выбросы.

### Анализ качественных переменных.

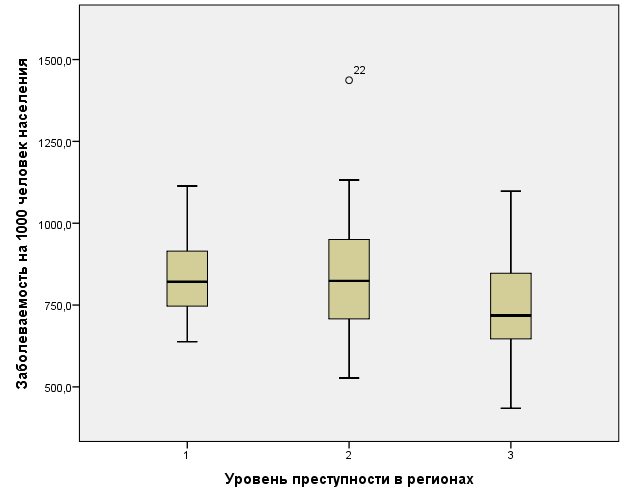
Здесь следует привести столбчатые диаграммы, которые отражают количество измерений с разными уровнями для данной переменной.



## Анализ статистической связи.

### Графический анализ пары «числовая зависимая переменная – качественная независимая переменная».

Здесь для каждой пары {количественная зависимая переменная – качественная независимая переменная} необходимо построить категоризированную диаграмму Бокса-Уискера (Box-Whisker).



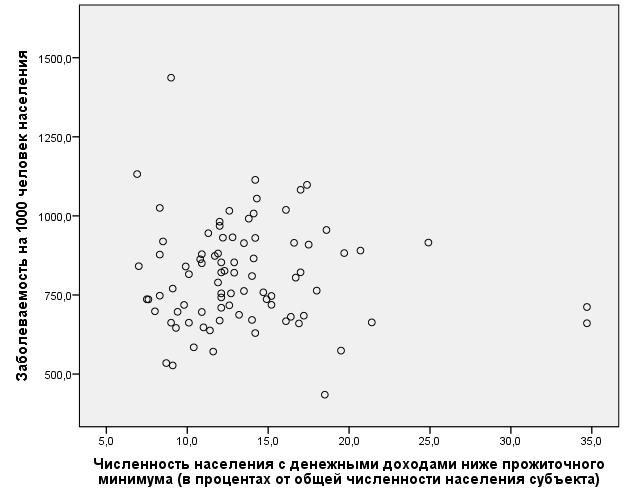
**Рисунок 6**

На основании анализа диаграммы следует охарактеризовать связь среднего значения и разброса количественной зависимой переменной с уровнями качественной независимой переменной.

Для формальной проверки гипотезы о наличии статистической связи следует выполнить непарамтерический дисперсионный анализ.

### Графический анализ пары «числовая зависимая переменная – числовая независимая переменная».

Здесь для каждой пары {количественная зависимая переменная – количественная независимая переменная} необходимо построить диаграммы рассеивания (Scatter plot).



**Рисунок 7**

На основании визуального анализа диаграммы следует сделать предположение о наличии и характере статистической взаимосвязи.

Для формальной проверки гипотезы о наличии связи следует подсчитать коэффициенты корреляции Пирсона и Спирмена, а также тау Кендала и привести результаты проверки их значимости.

### Анализ наличия корреляции между независимыми переменными.

Следует проанализировать силу связи между независимыми переменными, используя инструменты пп. 3.2.1 и 3.2.2. Для анализа силы связи между качественными переменными следует использовать анализ таблиц: необходимо привести таблицу кросс-табуляции, значения статистики хи-квадрат и V-Крамера.

### Предварительная проверка гипотез

Здесь необходимо рассказать о результатах проверки гипотез из п.2.3 на основании предварительного анализа данных.

# Спецификация, оценивание и оптимизация модели

На основании выдвинутых в п. 2.3. гипотез (альтернативные части гипотез), а также результатов предварительного анализа данных следует выписать функциональный вид детерминированной части модели линейной регрессии, которая будет использована для решения поставленной в п. 1.1 задачи. В состав модели можно включить все выбранные независимые переменные. Однако, при необходимости (слишком много коэффициентов) переменные, которые, судя по предварительному анализу не влияют на зависимую переменную, могут быть предварительно исключены. Для количественных переменных, влияние которых, предположительно, нелинейное, в модель включается линейная и нелинейная составляющая (квадрат переменной, гипербола, корень и т.д.). Для качественных переменных в модель включается соответствующий набор фиктивных переменных. Для проверки гипотезы о взаимодействии переменных в модель включается соответствующее слагаемое. Если предполагается взаимодействие между количественной и качественной переменной, в модель включаются соответствующие слагаемые (произведения фиктивных переменных и количественной переменной). Общее количество коэффициентов не должно превышать количество измерений, деленное на 30.

Следует не только выписать выражение для регрессии, но и пояснить свое решение.

## Связь структуры базовой модели с основными гипотезами исследования.

Здесь следует указать, как планируется проверять выдвинутые выше гипотезы, т.е. переписать их в виде ограничений на коэффициенты и сформулировать пары статистических гипотез.

## Оценивание базовой модели и результаты проверки гипотез.

Для дальнейшей проверки прогностических свойств модели следует разделить выборку на две части случайным образом в соотношении 80:20. Большая часть выборки (обучающая) будет использоваться для построения модели, меньшая (тестовая) — для проверки прогностических свойств.

Здесь следует привести результаты МНК оценивания в формате:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Переменная | Оценка коэффициента | Стандартная ошибка | t – статистика | p – уровень | 95% дов. интервал левая граница | 95% дов. интервал правая граница | VIF |
|  |  |  |  |  |  |  |  |

В этом разделе также следует выполнить проверку гипотез, сформулированных в 2.3 и 4.1 гипотез и дать интерпретацию результатам проверки. Также следует дать интерпретацию значениям коэффициента VIF и сопоставить полученные значения с результатами п. 3.2.3

## Анализ наличия выбросов

В этом разделе следует представить графики расстояния Кука и стьюдентизированных удаленных остатков и дать их интерпретацию. Используя данные инструменты, следует указать наблюдения, которые можно рассматривать, как выбросы и сформулировать, чем они отличаются от основной массы наблюдений.

## Анализ наличия гетероскедастичности

В данном разделе следует привести результаты проверки наличия гетероскедастичности с помощью критерия Уайта без учета взаимодействия.

## Оптимизация модели

В данном разделе следует привести результаты оптимизации модели на основе информационного критерия Акаике. Результаты представляются в виде таблицы результатов оценивания наилучшей модели в формате п. 4.2

## Проверка прогностических свойств модели

Для тестовой выборки построить прогнозы значений зависимой переменной и подсчитать среднеквадратическую погрешность. Построить множество 95% доверительных интервалов для значений зависимой переменной и подсчитать долю прогнозов, которые накрываются ими.

# Выводы и рекомендации

В данном разделе следует привести решение поставленной в разделе 2 задачи. Следует отметить наиболее яркие особенности собранных данных и результаты проверки гипотез. Необходимо рассказать, как потенциальный потребитель из п 2.2. может решать поставленную задачу с использованием построенной модели и опираясь на результаты проверки гипотез.