МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Дальневосточный федеральный университет»
(ДВФУ)

ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ
Департамент информационных и компьютерных систем

Курс «Компьютерные методы анализа больших данных»

Лабораторная работа №2
Контрольное мероприятие по рейтингу

на тему «Построение модели линейной регрессии для прогноза на основе датасета rand5»

Вариант №15

Выполнил студент Б9122-01.03.02мкт
Пелагеев Д.И.
_____

Проверил доцент Достовалов В.Н.

г. Владивосток
2024

Оглавление

Введение

Данный отчет посвящен процессу предварительного анализа данных из набора «Diamonds» в формате CSV. Нас интересуют следующие параметры:

— вес алмаза (в каратах);
— качество огранки;
— цвет алмаза;
— цена (в долларах США).

Актуальность данной работы обусловлена тем, что она служит хорошим материалом для изучения основных функций анализа данных в R.

Цель исследования — изучить возможности языка R для анализа данных и создать визуализации, которые позволят наглядно рассмотреть распределение характеристик алмазов.

Задачи работы:

1) Прочитать файл с помощью команды: `file <- read.csv(file="rand5.csv header=TRUE, sep= ")`

2) Для двух наихудших классов `color` и `cut` построить `subset`.

3) Построить модель линейной регрессии для построенного в п.2 `subset`. Объясняемая переменная - цена алмаза. Регрессор – вес алмаза.Сделать следующие виды прогнозов: точечный прогноз, прогноз с помощью предиктивного интервала, прогноз с помощью доверительного интервала. Прогноз построить для трехзначений веса, максимальное значение веса алмаза в subset увеличить на 5%, 10%, 15 % соответственно. Построить визуализацию уравнения регрессии с учетом доверительного интервала коэффициентов регрессии.

1 Подготовка данных

1.1 Загрузка и первичная обработка данных

Прежде чем мы начнем работать с данными в RStudio[1], стоит сначала установить рабочую директорию, в которой будет находиться наш датасет “rand5.csv“.

Листинг 1 — Установка директории

```
1    setwd("/Users/daniil/Desktop/Education/BigData")
```

Теперь загружаем датасет “rand5.csv“ с помощью функции `read.csv`.

Листинг 2 — Загрузка данных

```
1    file ← read.csv(file="rand5.csv", header=TRUE, sep=",")
```

Проверяем, существуют ли “NA“ элементы в нашем датасете с помощью функции `anyNA`. Если такие данные есть, то данная функция выведет “TRUE“, иначе “FALSE“.

Листинг 3 — Очистка данных

```
1    anyNA(diamonds)
```

1.2 Описание данных

Изучаем, какие данные находятся в файле.

Листинг 4 — Проверка данных

```
1    head(diamonds)
```

Датасет diamonds представляет собой набор данных, содержащий информацию о различных алмазах, собранную для анализа их характеристик и цен. Он включает 53 940 записей и 10 переменных, каждая из которых описывает определенные свойства алмаза[2]. У нас модифицированный датасет, который имеет 5 000 записей и 6 переменных:

— **X**: идентификатор строки;

— **carat**: масса алмаза в каратах;

— **cut**: качество огранки алмаза (Fair, Good, Very Good, Premium, Ideal);

— **color**: цвет алмаза, обозначаемый буквами от D (бесцветный) до J (с легким желтым оттенком);

— **clarity**: чистота алмаза, описываемая различными категориями (например, I1, SI1, SI2, VS1, VS2, VVS1, VVS2, IF);

— **price**: цена алмаза в долларах США.

Таблица 1 — Пример данных алмазов

X	carat	cut	color	clarity	price
1	1.20	Ideal	D	SI2	6140
2	0.37	Ideal	G	IF	1056
3	0.80	Ideal	F	VS2	3913
4	1.07	Ideal	H	SI1	4955
5	0.52	Very Good	F	VS2	1581
6	1.01	Ideal	E	SI2	4666

1.3 Формирование подмножеств

После всех манипуляций с датасетом мы должны отобрать два наихудших класса `color` и `cut`. С помощью `unique` мы составим массив только из уникальных значений в столбце, а далее с помощью функции `sort` отсортируем их в порядке их значимости. В итоге мы получим два наихудших элемента, которые поместим в новые переменные для дальнейшей работы.

Листинг 5 — Наихудшие цвета и огранки

```
1  sort(unique(diamonds$color))
2
3  worst_colors ← sort(unique(diamonds$color))[1:2]
4  worst_colors
5
6  sort(unique(diamonds$cut))
7
8  worst_cuts ← sort(unique(diamonds$cut))[1:2]
```

```
9       worst_cuts
```

Теперь, зная наихудшие цвета и огранки, мы можем выделить подвыборку на их основе. При этом проверим на существование в этой подвыборке “NA“, далее выведем первые строки для проверки корректности фильтрации.

Листинг 6 — Подмножество

```
1       subset ← diamonds[(diamonds$cut %in% worst_cuts) &
          (diamonds$color %in% worst_colors), ]
2
3       anyNA(subset)
4
5       head(subset)
```

Таблица 2 — Пример данных подвыборки

X	carat	cut	color	clarity	price
11	2.01	Good	D	SI2	16776
12	0.31	Good	E	SI1	544
84	0.32	Good	D	SI1	706
110	0.50	Good	E	VS2	1348
124	0.70	Good	D	SI1	2512
141	0.90	Fair	E	SI2	3342

2 Построение модели линейной регрессии

2.1 Теоретическая часть линейной регрессии

Теоретическая формула:

$$f(\mathbf{x}, \beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

- $\mathbf{x} = (x_1, x_2, \ldots, x_n)$: Вектор независимых переменных.

- $\beta = (\beta_0, \beta_1, \ldots, \beta_n)$: Вектор коэффициентов модели.

- β_0: Свободный член (интерцепт).

- $\beta_1, \beta_2, \ldots, \beta_n$: Коэффициенты при соответствующих независимых переменных.

- ϵ: Случайная ошибка, отражающая отклонение наблюдаемого значения от предсказанного моделью.

Формула для нашего случая:

$$\text{price}_i = \beta_0 + \beta_1 \cdot \text{carat}_i + \epsilon_i$$

- carat_i: Вес бриллианта в каратах (независимая переменная).

- β_0: Предсказанная цена бриллианта при нулевом весе (хотя физически нулевой вес невозможен, этот параметр служит для определения смещения линии регрессии).

- β_1: Изменение предсказанной цены при увеличении веса бриллианта на один карат.

- ϵ_i: Случайная ошибка для i-го наблюдения, отражающая отклонение фактической цены от предсказанной моделью.

2.2 Создание модели линейной регрессии

Функция `lm` создает модель линейной регрессии, где зависимая переменная `price`, а независимая переменная `carat`. Все данные мы берем из построенной нами подвыборки “subset“.

Листинг 7 — Создание модели

```
1    lm_model ← lm(price ~ carat, data = subset)
```

Благодаря полученной модели мы сможем чуть позже узнать предикативный и доверительный интервалы, а также выполнить точечный прогноз.

2.3 Получение сводки модели

Чтобы получить подробную информацию о модели, мы используем функцию `summary`. Сводка модели предоставляет информацию о коэффициентах, их значимости, качестве подгонки модели и других статистических показателях.

Листинг 8 — Сводка модели

```
1    s ← summary(lm_model)
```

```
2 s
```

Вывод этого блока выглядит так:

```
    Call:
lm(formula = price ~ carat, data = subset)

Residuals:
    Min      1Q  Median      3Q     Max
-9035.2  -666.4   -97.3   433.7  8452.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1862.1      134.3  -13.87   <2e-16 ***
carat         6532.1      137.7   47.42   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1553 on 585 degrees of freedom
Multiple R-squared:  0.7936, Adjusted R-squared:  0.7932
F-statistic:  2249 on 1 and 585 DF,  p-value: < 2.2e-16
```

2.4 Анализ линейной регрессии

— **(Intercept)**: -1574.0 — значение цены при нулевом весе (теоретически, но в реальности физически невозможный).

— **carat**: 6433.4 — увеличение веса на 1 карат связано с увеличением цены на 6433.4 USD.

— Оба коэффициента имеют $Pr(> |t|) < 2 \cdot 10^{-16}$, что указывает на их высокую статистическую значимость.

— **Multiple R-squared**: 0.8475 — модель объясняет примерно 84.75% вариации цены.

— Высокое значение и низкое p-value ($< 2.2 \cdot 10^{-16}$) подтверждают, что модель значимо лучше, чем модель без предиктора.

Из этих данных видно, что модель показывает сильную линейную зависимость между весом и ценой бриллианта в выбранных категориях. Высокий R-квадрат указывает на хорошую подгонку модели.

3 Прогнозирование цен для увеличенных значений веса алмазов

3.1 Определение новых значений `carat` для прогнозирования

С помощью функции `max` мы определим максимальное значение столбца “carat“ в датасете “diamonds“. Далее мы в соответствии с заданием увеличим этот вес на 5%, 10% и 15%. И эти новые значения поместим во вектор. После создадим датафрейм из нового вектора для дальнейшей работы.

Листинг 9 — Максимальные значения carat

```
max_carat ← max(subset$carat)
max_carat

max_carats ← data.frame(carat = max_carat * c(1.05, 1.10,
    1.15))
max_carats

new_data ← data.frame(carat = max_carats)
new_data
```

Таблица 3 — Значения увеличенного веса алмазов

	carat
1	3.57
2	3.74
3	3.91

3.2 Выполнение прогнозов и составление датафрейма

Блок ниже поможет нам составить точечное прогнозирование цены, предиктивные интервалы и доверительные интервалы для новых зна-

чений “carat“. В этом нам поможет функция `predict`, только с разными интервалами. Если не указываем интервал, то получим точечное прогнозирование, указываем “prediction“ для предикативного интервала и “confidence“ для доверительного интервала. Далее нужные значения помещаем в датафрейм и выводим его.

Листинг 10 — Прогнозы и интервалы

```
point_predictions ← predict(lm_model, newdata = new_data)

predictions_pred ← predict(lm_model, newdata = new_data,
                           interval = "prediction", level =
                              0.95)[, -1]
predictions_conf ← predict(lm_model, newdata = new_data,
                           interval = "confidence", level =
                              0.95)[, -1]

results_table ← data.frame(
  Carat = new_data$carat,
  Point_Pred = point_predictions,
  Pred_Lwr = predictions_pred[, "lwr"],
  Pred_Upr = predictions_pred[, "upr"],
  Conf_Lwr = predictions_conf[, "lwr"],
  Conf_Upr = predictions_conf[, "upr"]
)
results_table
```

Таблица 4 — Результаты прогнозирования стоимости алмазов

ID	Carat	Point_Pred	Pred_Lwr	Pred_Upr	Conf_Lwr	Conf_Upr
1	3.57	21457.52	18317.42	24597.62	20712.70	22202.33
2	3.74	22567.98	19416.81	25719.14	21777.80	23358.16
3	3.91	23678.44	20515.57	26841.30	22842.82	24514.05

4 Подготовка данных для визуализации

4.1 Подготовка данных для прогноза и графиков

Помещаем информацию о весах алмазов в переменную и датафрейм для дальнейших действий.

Листинг 11 — Подготовка данных для графиков

```
carat_range ← subset$carat
head(carat_range)

plot_data ← data.frame(carat = carat_range)

head(plot_data)
```

4.2 Получение прогнозов и объединение их

Сделаем новые переменные для визуализации предективного и доверительного интервала. Данные беруться из нашей подвыблорки.

Листинг 12 — Получение и объединение прогнозов

```

visual_predictions_conf ← predict(lm_model, newdata =
   plot_data, interval = "confidence", level = 0.95)
visual_predictions_conf ←
   as.data.frame(visual_predictions_conf)
visual_predictions_conf$carat ← carat_range
visual_predictions_conf$IntervalType ← "Confidence interval"

visual_predictions_pred ← predict(lm_model, newdata =
   plot_data, interval = "prediction", level = 0.95)
visual_predictions_pred ←
   as.data.frame(visual_predictions_pred)
visual_predictions_pred$carat ← carat_range
visual_predictions_pred$IntervalType ← "Predictive interval"

combined_predictions ← rbind(
  visual_predictions_conf[, c("carat", "lwr", "upr",
     "IntervalType")],
  visual_predictions_pred[, c("carat", "lwr", "upr",
     "IntervalType")]
)
```

5 Визуализация данных

Визуализацию модифицированного набора diamonds будем производить с помощью библиотеки ggplot2 [3], которая позволяет создавать наглядные графики на языке R[4].

5.1 График линейной реграссии

График был построен так, чтобы визуализировать модель линейной регрессии с доверительными и предиктивными интервалами. Каждые элементы графика были описаны разными цветами для лучшего восприятия. С этой же целью были добавлены заголовки, подписи осей и легенды, поясняющей типы интервалов. Такой подход позволяет визуально оценить качество модели и её предсказаний.

Листинг 13 — Код для гистограммы

```
ggplot(subset, aes(x = carat, y = price)) +
  geom_point(color = "blue") +
  geom_line(data = visual_predictions_conf, aes(x = carat, y =
      fit), color = "red") +
  geom_ribbon(data = combined_predictions,
              aes(x = carat, ymin = lwr, ymax = upr, fill =
                  IntervalType),
              alpha = 0.2, inherit.aes = FALSE) +
  labs(title = "Linear regression model: Price vs Weight",
       subtitle = "With confidence and predictive intervals",
       x = "Weight (carat)",
       y = "Price (USD)",
       fill = "Interval type") +
  scale_fill_manual(values = c("Confidence interval" = "green",
                               "Predictive interval" =
                                 "orange")) +
  theme_minimal() +
  theme(legend.position = "right")
```

На графике видно, что существует четкая положительная зависимость между весом и ценой, так как точки сгруппированы вдоль линии регрессии. Ширина как доверительного, так и предиктивного интервала увеличивается с ростом веса, что может быть связано с меньшим количе-

ством наблюдений для более тяжелых алмазов и увеличением вариации цены. Но в общем и целом видно, что модель линейной регрессии достаточно хорошо описывает данные в этом подмножестве, а включенные интервалы дают полезную информацию о достоверности прогнозов.

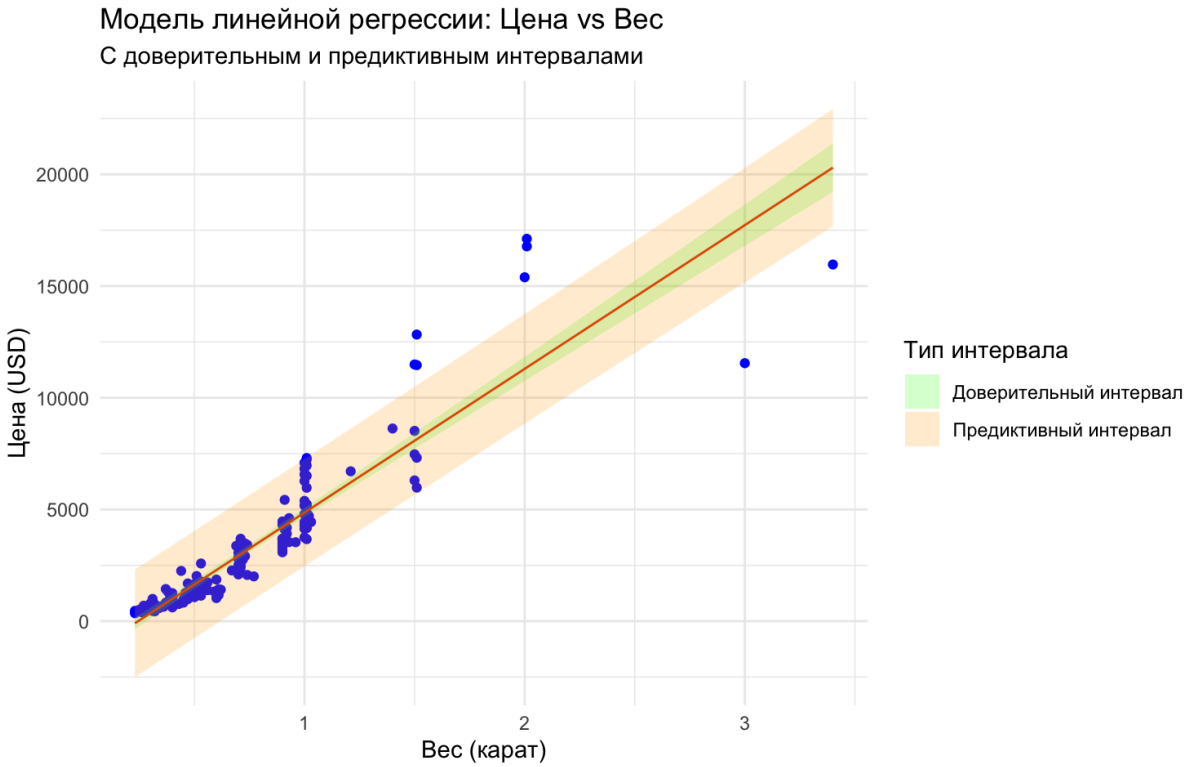

Рисунок 1 — График модели линейной регрессии с доверительным и предиктивным интервалами

Заключение

В ходе проведённого анализа мы изучили зависимость цены алмаза `price` от его веса `carat` на основе очищенных данных. Данные были загружены, очищены от пропусков и отфильтрованы для анализа двух наименее благоприятных категорий по параметрам `cut` и `color`. С помощью линейной регрессии была создана модель, демонстрирующая значимое влияние веса алмаза на его цену.

На основе модели были сделаны прогнозы цен для увеличенных значений веса алмазов, включая доверительные и предиктивные интервалы, что позволило оценить надёжность и вариабельность прогнозов. Визуализация результатов с использованием `ggplot2` наглядно отразила исходные данные, линию регрессии, а также доверительные и предиктивные интервалы, обеспечивая чёткое понимание зависимости между весом и ценой алмазов.

Полученные результаты подтверждают, что вес алмаза является значимым фактором, влияющим на его цену. Модель линейной регрессии эффективно описывает эту зависимость и позволяет делать обоснованные прогнозы.

Список использованных источников

1. Официальная страница интегрированной среды разработки RStudio. — 2024. — Режим доступа: RStudio | Open source & professional software for data science teams - RStudio дата обращения: 24.11.2024).

2. *R Foundation for Statistical Computing.* Diamonds dataset. — 2023. — Режим доступа: Diamonds dataset (дата обращения: 24.11.2024).

3. *Wickham, Hadley.* ggplot2: Elegant Graphics for Data Analysis (Use R!) / Hadley Wickham. — New York: Springer, 2009. — Режим доступа: ggplot2 (дата обращения: 24.11.2024).

4. Официальная страница среды статистического моделирования R. — 2024. — Режим доступа: R: The R Project for Statistical Computing (дата обращения: 24.11.2024).