УДК 51-7

РЕГРЕССИОННАЯ МОДЕЛЬ ЗАВИСИМОСТИ АРТЕРИАЛЬНОГО ДАВЛЕНИЯ ОТ ВОЗРАСТА

Гребенюк Б.Л., Тимергалиев А.И.

Уфимский колледж радиоэлектроники, телекоммуникаций и безопасности, г. Уфа, Россия

Научный руководитель: Мухаметрахимова А.И. – преподаватель математики УКРТБ

Аннотация. В данной работе с помощью методов регрессионного анализа исследуется зависимость артериального давления от возраста. Для проведения данного исследования нами были сняты показания артериального давления у 100 испытуемых. Построение регрессионной модели было проведено в среде программирования R.

Ключевые слова: регрессия, регрессионная модель, нормальное распределение, математическое ожидание, среднеквадратичное отклонение.

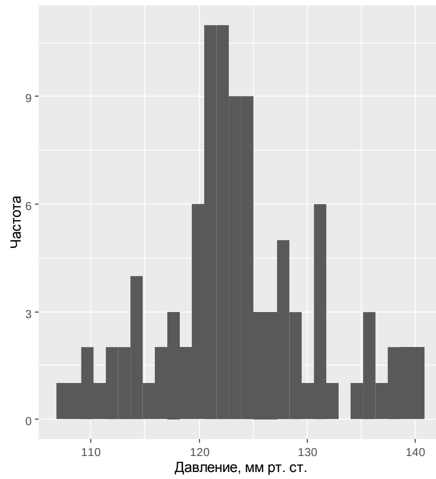
Введение. В настоящее время методы регрессионного анализа нашли широкое применение в различных областях науки при исследовании различных процессов и явлений. Можно выделить несколько основных причин, объясняющих широкое применение данного метода. Во-первых, регрессионный анализ позволяет сделать статистические выводы и объяснить некоторые научные феномены. Во-вторых, регрессионный анализ удобен тем, что представляет собой компактное описание зависимостей исследуемых параметров системы. В-третьих, построение регрессионной модели для выборочных наблюдений позволяет сделать обобщение и распространение статистических выводов на всю генеральную совокупность. В-четвертых, построение регрессионной модели помогает прогнозировать ожидаемые значения зависимой переменной.

Цель работы: построить регрессионную модель зависимости артериального давления от возраста.

Основная часть. Вначале нами были сняты показания артериального давления среди 100 человек, в состав которых вошли студенты и сотрудники ГБПОУ УКРТБ, а также знакомые и родственники. На первом этапе нашего исследования при проведении измерений показателей артериального давления мы не учитывали возраст людей. Для достижения поставленной цели мы использовали статистический пакет R [1]. Ниже представлены полученные нами значения артериального давления.

```
> y <- c(
+ 112.14, 117.55, 110.76, 115.26, 117.13, 120.39, 121.03,
+ 114.03, 124.83, 113.92, 122.04, 113.41, 123.61, 130.93,
+ 116.64, 117.06, 111.73, 120.41, 112.98, 117.20, 120.19,
+ 128.53, 120.14, 108.70, 130.77, 110.16, 129.07, 123.46,
+ 130.02, 123.31, 124.06, 129.17, 131.08, 107.62, 139.77,
+ 121.47, 130.95, 128.15, 114.31, 134.58, 125.86, 138.49,
+ 110.01, 127.80, 122.57, 136.99, 139.53, 127.34, 132.26,
+ 120.85, 122.99, 121.36, 122.46, 123.58, 123.05, 127.83,
+ 140.42, 122.64, 121.01, 135.69, 138.25, 127.24, 135.55,
+ 124.76, 122.67, 123.33, 127.00, 125.00, 123.98, 113.81,
+ 122.92, 123.97, 121.45, 130.68, 126.98, 121.45, 120.22,
+ 124.10, 121.60, 118.79, 120.93, 123.85, 122.28, 121.97,
+ 135.59, 119.62, 139.12, 125.96, 124.94, 118.87, 127.64,
+ 120.64, 122.46, 121.03, 122.56, 123.31, 121.56, 130.87,
+ 122.52, 123.97)
> c(mean(y), sd(y))
[1] 123.647300  7.258169
> shapiro.test(y)

        Shapiro-Wilk normality test

data:  y
W = 0.97011, p-value = 0.02255

> library(ggplot2)
> ggplot(data = data.frame(y), aes(x = y)) + geom_histogram() +
+  ylab("Частота") + xlab("Давление, мм рт. ст.")
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

В результате получили диаграмму, представленную ниже. Проанализировав вид полученной диаграммы, приходим к выводу, что мы имеем дело с нормальным распределением переменной.



Нормальное распределение характеризуется двумя параметрами: математическим ожиданием μ и стандартным отклонением σ. Так как распределение исследуемой случайной величины Y задается с помощью нормального распределения, то статистическую модель можем представить в виде

$$y_i \sim N(\mu, \sigma).$$

В нашем случае выборочное среднее значение и стандартное отклонение представляют 123,65 мм.рт.ст. и 7,26 мм.рт.ст. соответственно. Следовательно, нашу модель можем представить в виде:

$$y_i = 123{,}65 + \varepsilon_i,$$

где через ε_i обозначено остатки модели.

Проверим насколько хорошо параметры артериального давления, которые были сгенерированы простейшей статистической моделью согласуются с экспериментальными данными. С помощью рассмотренной выше простой моделью мы можем легко сгенерировать в R новые параметры.

```
> set.seed(101)
> y.new.1 <- rnorm(n = 100, mean = 123.65, sd = 7.26)
> set.seed(101)
> y.new.2 <- 123.65 + rnorm(n = 100, mean = 0, sd = 7.26)
> all(y.new.1 == y.new.2)
[1] TRUE
```

Следует отметить, что характеристики рассмотренной простейшей статистической модели не являются точным, а представляют лишь точечные оценки истинных параметров. И поэтому всегда присутствует некая неопределенность в отношении того насколько истины эти показатели.

Для проведении имитаций воспользуемся функцией lm().

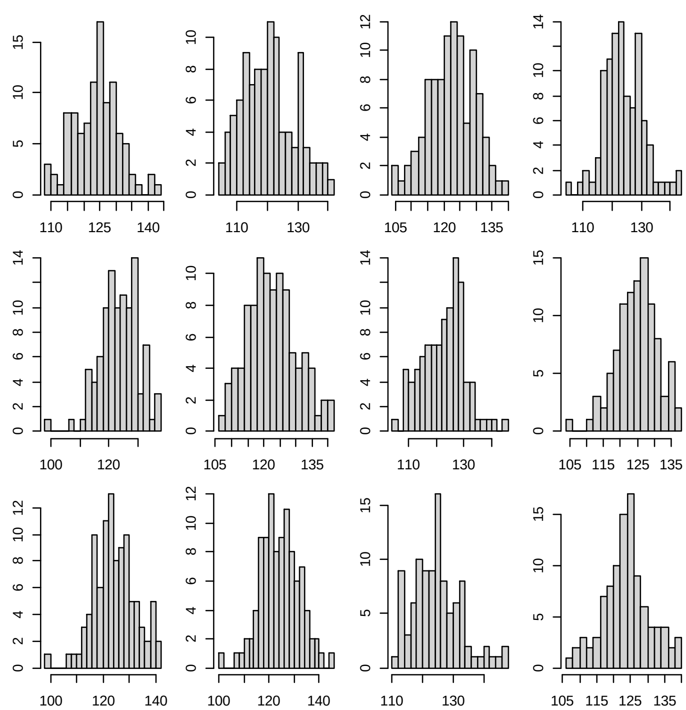
```
Call:
lm(formula = y ~ 1)

Residuals:
     Min       1Q   Median       3Q      Max
-16.0273  -3.2998  -0.6923   4.0327  16.7727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 123.6473     0.7258   170.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.258 on 99 degrees of freedom
```

Из полученных результатов мы видим, что свободный член подогнанной модели в точности совпадает со средним значением (123, 65 мм. рт. ст.), а стандартное отклонение остатков модели совпадает со стандартным отклонением этих данных (7.258 мм. рт. ст.). Мы также вычислили оценку стандартной ошибки среднего значения, равную 0,7258. Стандартная ошибка – стандартное отклонение нормального распределения значений данного параметра, рассчитанных по выборкам одинакового размера из той же генеральной совокупности. Зная данные параметры, мы можем сгенерировать новые данные. Решим эту задачу с использованием функции sim().

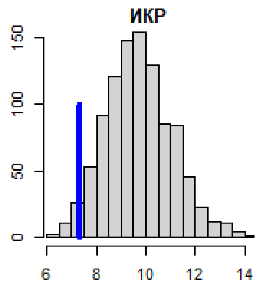
Сгенерируем 1000 альтернативных реализаций среднего кровяного давления.

```
> set.seed(102)
> y.sim <- sim(y.lm, 1000)
> y.rep <- array(NA, c(1000, 100))
> for(s in 1:1000){
+  y.rep[s, ] <- rnorm(100, y.sim@coef[s], y.sim@sigma[s])
+  }
```

Изобразим гистограммы выборочных распредлений значений кровяного давления, сгенерированных на основе первых 12 реализаций простой статистической модели.



Рассчитаем интерквартильный размах (ИКР) для каждого имитированного набора данных и сравним полученное распределение с 1000 таких значений с ИКР реальных данных. Ниже представим гистограмму 1000 значений ИКР, рассчитанных для каждого имитированных распределений кровяного давления. Вертикальной синей линией покажем ИКР для реально наблюдаемых значений.

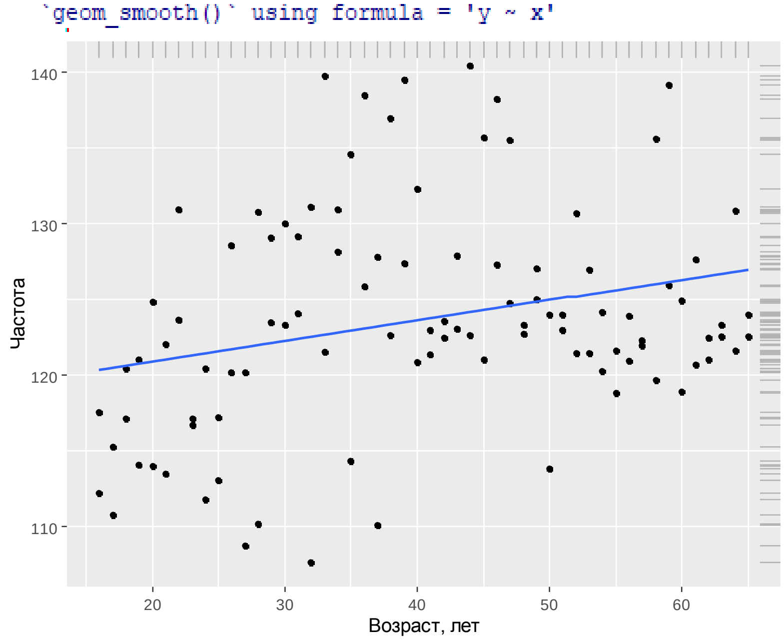


На полученной диаграмме мы хорошо видим, что значения ИКР для имитированных данных занижены по сравнению реальными данными. А это означает, что простейшая статистическая модель недооценивает уровень вариации реальных значений кровяного давления. Возможная причина такого несоответствия заключается в том, что мы не учитываем влияние на показатели кровяного давления различных факторов. Например, возраст, пол, масса, состояние здоровья и т.д.

Теперь исследуем зависимость показателя артериального давления от возраста. То есть помимо фиксирования показателей артериального давления, мы также учитывали и возраст каждого испытуемого.

Ниже представлена графическая связь между возрастом и значением артериального давления. Для визуализации тренда в данных добавим линию регрессии. Она изображена с помощью линии синего цвета.

```
> x <- rep(seq(16, 65, 1), each = 2)
> Data <- data.frame(Age = x, BP = y)
> ggplot(data = Data, aes(x = Age, BP)) + geom_point() +
+  geom_smooth(method = "lm", se = FALSE) +
+  geom_rug(color = "gray70", sides = "tr") +
+  ylab("Частота") + xlab("Возраст, лет")
`geom_smooth()` using formula = 'y ~ x'
```



Из графика видим, что между показателями артериального давления и значением возраста есть линейная зависимость. При увеличении возраста увеличивается и показания давления. Учесть данное изменение среднего давления артериального давления мы можем, добавив возраст в простейшую статистическую модель:

$$y_i \sim N(\beta_0 + \beta_1 \cdot Age_i + \varepsilon_i, \sigma).$$

Параметры β_0 и β_1 вычислим с помощью метода наименьших квадратов. По полученным данным модель кровяного давления можно представить в виде:

$$y_i \sim N(118.173 + 0.136 \cdot Age_i + \varepsilon_i, \sigma),$$

где $\varepsilon_i \sim N(0{,}7.712)$.

Графически данная модель изображена на графике синей линией.

Таким образом, статистическая модель, учитывающая возраст, описывает более точно вариацию значений кровяного давления у 100 наших испытуемых, чем простейшая статистическая модель.

Заключение. В данной работе исследуется зависимость артериального давления от возраста человека. Для проведения данного исследования нами были сняты показания артериального давления у 100 испытуемых, в состав которых вошли студенты и сотрудники ГБПОУ УКРТБ, знакомые и родственники. Зависимость давления от возраста мы исследовали с помощью построения регрессионной модели в среде программирования R. Проведя данное исследование, мы пришли к выводу, что между показателями давления и возраста человека существует линейная зависимость. С увеличением возраста человека показатели артериального давления также увеличиваются. Таким образом наше исследование подтвердило, что в зоне риска развития артериальной гипертензии находятся люди старшего возраста.

Список литературы

1. *Венэбльз У. Н., Смит Д. М. Введение в R. Заметки по R: среда программирования для анализа данных и графики // Москва. – 2013. – 109 с.*
2. *Демиденко Е. З. Линейная и нелинейная регрессии // Москва: Финансы и статистика. – 2018. –302 с.*

UDC 51-7

REGRESSION MODEL OF THE DEPENDENCE OF BLOOD PRESSURE ON AGE

Grebenyuk B.L., Timergaliev A.I.

Ufa College of Radio Electronics, Telecommunications and Security, Ufa, Russia

Mukhametrakhimova A.I. – teacher of mathematics at UKRTB

Annotation. In this paper, using regression analysis methods, the dependence of blood pressure on age is investigated. To conduct this study, we took blood pressure readings from 100 subjects. The regression model was built in the R programming environment.

Keywords. regression, regression model, normal distribution, mathematical expectation, standard deviation.