

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ ДНР  
**ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ**

**ФИЗИКО-ТЕХНИЧЕСКИЙ ФАКУЛЬТЕТ**

**ЛАБОРАТОРНАЯ РАБОТА №1**

По теме: «Построение наивного байесовского классификатора»

По предмету: «Машинное обучение»

Выполнил:

студент 2 курса

группы ИВТ-3

Дворников Г.Н

Проверил:

Савенков И.Н.

Донецк  
2021

**Лабораторная работа № 1**

**Тема**: интеллектуальный анализ данных средствами языка R: обработка статистических данных и построение модели регрессии.

**Цель:** изучить приемы исследования корреляционной зависимости, построения парной и множественной линейной регрессии средствами языка R.

**Ход** **работы**

Загрузка и анализ данных

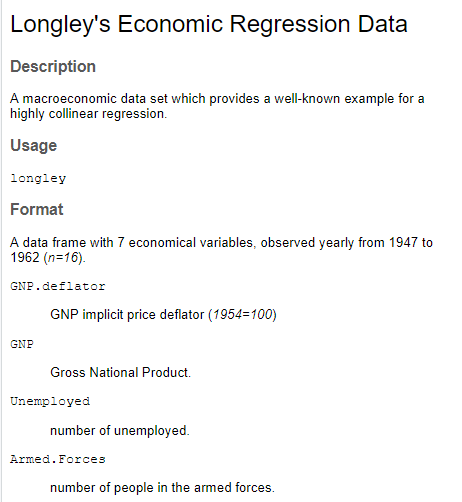


Рисунок 1 - Справка по набору данных longley

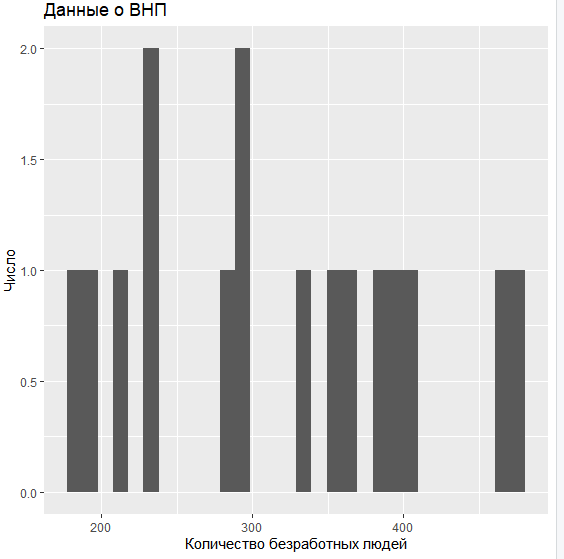


Рисунок 2 - Гистограмма абсолютных частот

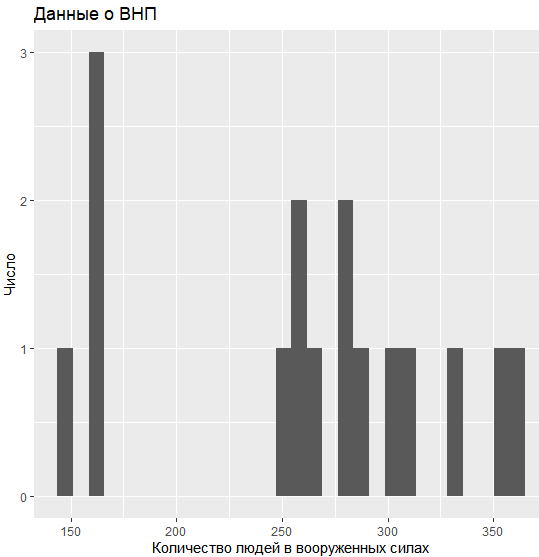


Рисунок 3 - Гистограмма абсолютных частот

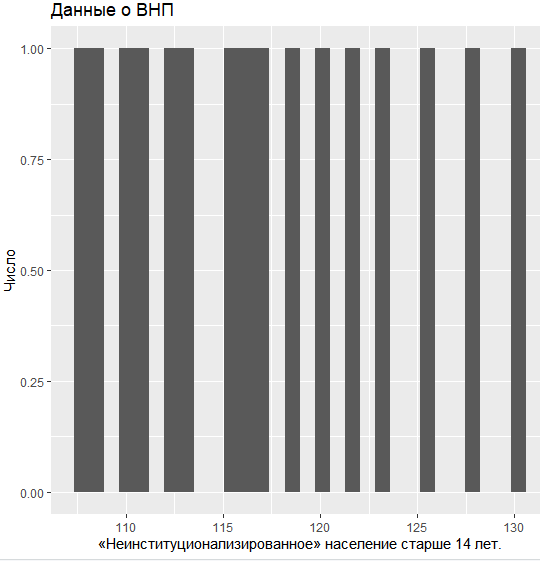


Рисунок 5 - Гистограмма абсолютных частот

Оценили модель линейной регрессии. Для этого командой lm поместили в переменную model модель линейной регрессии, указав dist в качестве зависимой переменной, и через значок ~ вторую переменную в качестве регрессора

Провели на графике полученную линию регрессии с 95% доверительными интервалами.

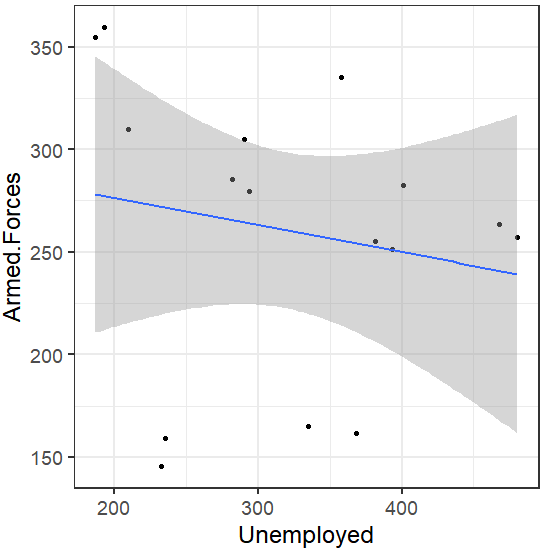


Рисунок 6 - График линейной регрессии с доверительным интервалом

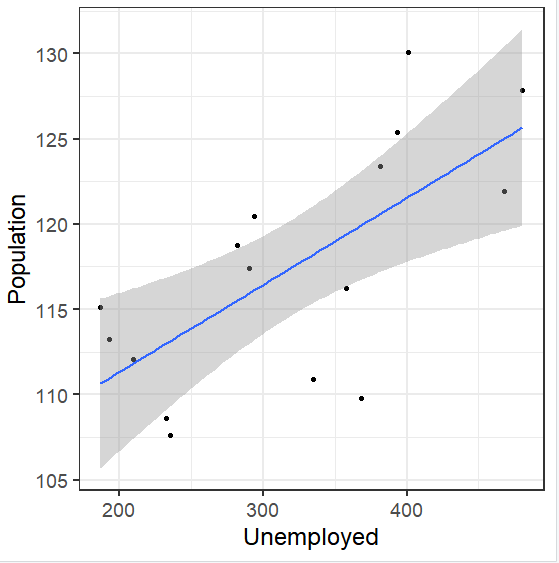


Рисунок 7 -График линейной регрессии с доверительным интервалом

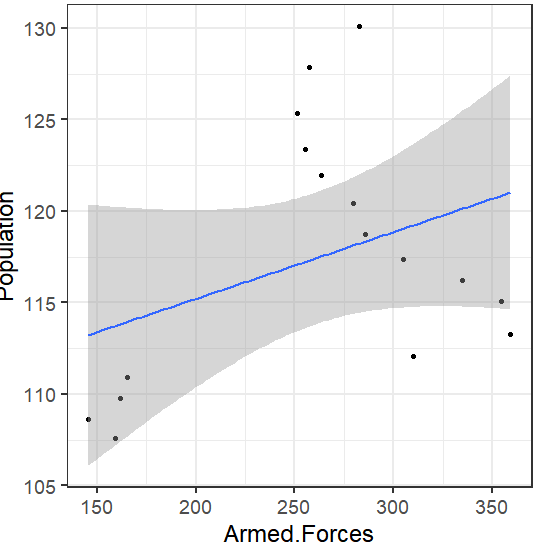


Рисунок 8 - График линейной регрессии с доверительным интервалом

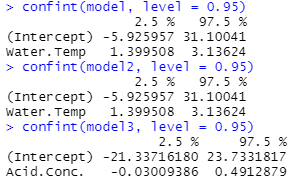


Рисунок 9 - 95% доверительные интервалы для параметров линейной регрессии

Построили модели множественной линейной регрессии.

Встроенным пакетом graphics содержащим функцию pairs, позволяющую получить все возможные диаграммы рассеяния на одном графике, а также выполнили их сглаживание с помощью опции panel.smooth.

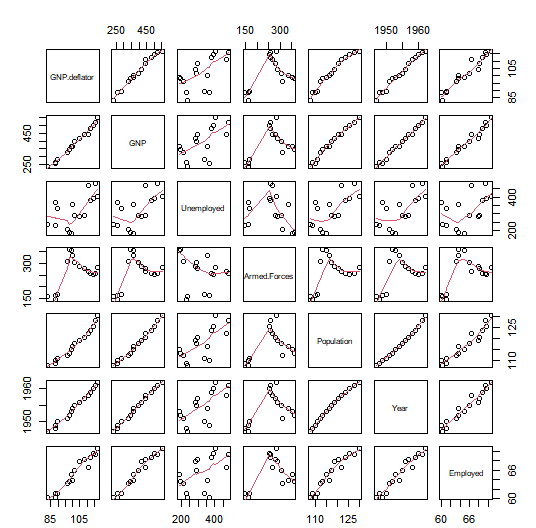


Рисунок 10 - Диаграммы рассеяния, полученные с помощью функции pairs

Функция cor вычислила корреляцию между двумя выборками, так и получили корреляционную матрицу для всех переменных из набора данных

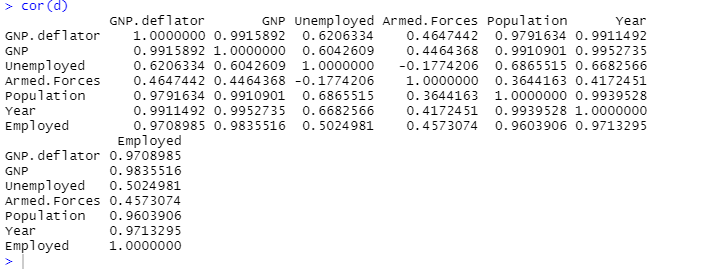


Рисунок 11 - Корреляционная матрица для всех переменных

Чтобы оценить регрессию рождаемости на остальные переменные, воспользовались уже знакомой функцией lm, а регрессоры перечислили через знак «плюс»

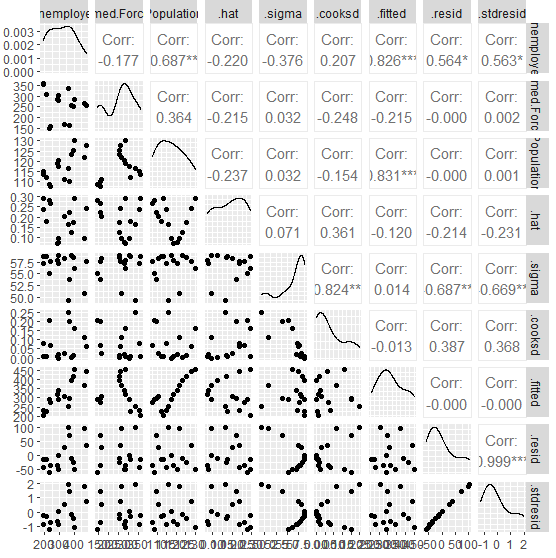


Рисунок 12 - Корреляционная матрица, диаграммы рассеяния и сглаженные распределения, полученные с помощью функции ggpairs

Получили оценки коэффициентов уравнения регрессии, а также проверили основные гипотезы.

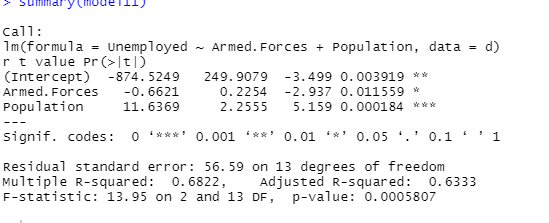


Рисунок 13 - Оценки коэффициентов уравнения регрессии

Построили прогноз по аналогии с парной линейной регрессией. Отличие заключается лишь в том, что в наборе данных указали значения каждого фактора

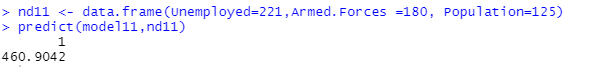


Рисунок 14 -Прогноз множественной линейной регрессии

**Вывод**: изучили приемы исследования корреляционной зависимости, построения парной и множественной линейной регрессии средствами языка R.

**Листинг скрипта**

data("longley")

d<-longley

qplot(data = d, Unemployed, xlab="Количество безработных людей", ylab="Число",main="Данные о ВНП")

qplot(data = d, Armed.Forces, xlab="Количество людей в вооруженных силах", ylab="Число",main="Данные о ВНП")

qplot(data = d, Population, xlab="«Неинституционализированное» население старше 14 лет.", ylab="Число",main="Данные о ВНП")

ggplot() +

+ geom\_point(aes(x=d$Unemployed, y=d$Armed.Forces), size = 2) + theme\_bw(base\_size =

+ 18) +

+ xlab("Количество безработных людей") + ylab("Количество людей в вооруженных силах") +

+ labs(title = "Корреляционное поле")

ggplot() +

+ geom\_point(aes(x=d$Unemployed, y=d$Population), size = 2) + theme\_bw(base\_size =

+ 18) +

+ xlab("Количество безработных людей") + ylab("«Неинституционализированное» население старше 14 лет") +

+ labs(title = "Корреляционное поле")

ggplot() +

+ geom\_point(aes(x=d$Armed.Forces, y=d$Population), size = 2) + theme\_bw(base\_size =

+ 18) +

+ xlab("Количество людей в вооруженных силах") + ylab("«Неинституционализированное» население старше 14 лет") +

+ labs(title = "Корреляционное поле")

model <- lm(data=d, Unemployed~Armed.Forces)

model$coefficients

model2 <- lm(data=d, Unemployed~Population)

model2$coefficients

model3 <- lm(data=d, Armed.Forces~Population)

model3$coefficients

qplot(data = d, Unemployed, Armed.Forces) + stat\_smooth(method="lm", level = 0.95) + theme\_bw(base\_size = 18)

`geom\_smooth()` using formula 'y ~ x'

qplot(data = d, Unemployed, Population) + stat\_smooth(method="lm", level = 0.95) + theme\_bw(base\_size = 18)

`geom\_smooth()` using formula 'y ~ x'

qplot(data = d, Armed.Forces, Population) + stat\_smooth(method="lm", level = 0.95) + theme\_bw(base\_size = 18)

`geom\_smooth()` using formula 'y ~ x'

RSS <- deviance(model)

RSS2 <- deviance(model2)

RSS3 <- deviance(model3)

TSS <- sum((stack.loss-mean(stack.loss))^2)

predict(model,nd)

predict(model2,nd2)

predict(model3,nd3)

pairs(d , panel = panel.smooth)

cor(d)

model11 <- lm(data=d, Unemployed~Armed.Forces+Population)

ggpairs(model11 , panel = panel.smooth)

summary(model11)

nd11 <- data.frame(Unemployed=221,Armed.Forces =180, Population=125)

predict(model11,nd11)

**Контрольные вопросы**

**1.** **Что понимается под регрессией случайной величины Y на X?**

В общем виде под регрессией случайной величины Y на X понимается зависимость, задающая траекторию движения точки (yx, x), которая определяется условным математическим ожиданием E(Y|X = x) для каждого текущего значения x.

**2. Могут ли зависимости в регрессионной модели быть нелинейными?**

Да

**3. Приведите модель линейной регрессии.**

y = b0 + b1 fj (x1,…, xq) +b2 fj (x1,…, xq) +…+bm fj (x1,…, xq),

где q – число независимых переменных, m – число регрессоров, включенных в модель.

**4. В чем заключается общий подход при нахождении оптимальных оценок параметров регрессионной модели?**

Общий подход при нахождении оптимальных оценок параметров модели заключается в минимизации некоторой выбранной функции потерь, учитывающей разности между прогнозируемыми и фактическими значениями отклика.

В качестве функции потерь используют сумму квадратов ошибок:

где ε ̂i – отклонения выборочных величин yi зависимой переменной от значений y ̂i, получаемых по уравнению регрессии, n – объем выборки. Если выборочные остатки модели регрессии распределены нормально, то метод наименьших квадратов имеет несомненное преимущество, поскольку позволяет получить несмещенные, наиболее эффективные и однозначно определенные оценки параметров.

**5. Какие** **критерии оценки качества регрессионной модели вы знаете?**

Коэффициент детерминации (R 2 ), скорректированный коэффициент детерминации (Adj R 2 ), статистика Фишера (F -статистика), нецентрированные коэффициенты детерминации.

**6. Что такое коллинеарность? Что такое мультиколлинеаность?**

Коллинеа́рность - отношение параллельности векторов: два ненулевых вектора называются коллинеарными, если они лежат на параллельных прямых или на одной прямой.

Мультиколлинеарность - наличие линейной зависимости между объясняющими переменными (факторами) регрессионной модели.

**7. Как решают вопрос о процедуре включения предикторов в модель регрессии в условиях мультиколлинеарности?**

Возможные варианты – выделение групп независимых переменных, проведение факторного анализа или другим вариантом отбора переменных при построении модели является метод пошагового включения переменных в регрессионную модель (Forward stepwise) или пошагового исключения переменных из модели (Backward stepwise).

**8. Что используют для контроля качества модели?**

Для оценки качества полученной модели используются критерии, при расчете которых учитывается число предикторов m:

стандартное отклонение для остатков

коэффициент детерминации

скорректированный коэффициент детерминации

который налагает штраф на добавление новых параметров в регрессионную модель;

F-статистика , р-значение для оценки статистической значимости которого получают с использованием распределения Фишера;

информационный критерий Акайке AIC=n ln(∑▒ε ̂\_i^2 /n)+2m и некоторые его разновидности – байесовский информационный критерий (BIC), критерий Шварца и др.

**9. Как проверяют модель на адекватность?**

В качестве проверки на адекватность построенной регрессионной модели анализируют выборку остатков полученной модели на близость к нормальному закону, поскольку именно выполнение предположений нормальности обеспечивает корректность выводов, связанных с гипотезами, рассматриваемыми в регрессионном анализе.

В случае, если проверка на качество и/или адекватность модели показала отрицательный результат, необходимо вернуться к началу и пересмотреть вид модели регрессии.

**10. Опишите схему действий при проведении регрессионного анализа.**

Рассмотрим примерную схему действий при проведении регрессионного анализа.

1. По рассматриваемым данным определить вид регрессионной модели: переменные, включаемые в модель в качестве регрессоров (отбор «вручную» с учетом корреляции xi, либо с использованием пошаговых процедур типа stepwise или процедур выбора регрессоров по всем подмножествам), их взаимодействие и характер влияния на значения отклика. При необходимости, преобразовать переменные в фиктивные (т.н. dummy переменные).

2. Построить регрессионную модель по выбранным независимым переменным, получить оценки параметров модели, значение коэффициента детерминации R2, проверить значимость регрессоров.

3. По значению R2 сделать вывод о значимости построенной модели: в случае, когда R2 > 0.75, т.е. включенные регрессоры объясняют более 75% изменчивости отклика, можно считать, что модель в достаточной степени отражает взаимосвязи между зависимой и независимыми переменными, и переходить к пункту 4. Иначе необходимо изменить структуру модели, в т.ч. включив в рассмотрение новые, не используемые ранее регрессоры от исходных переменных, и вернуться к пункту 1 .

4. Проверить адекватность полученной модели путем проверки остатков на нормальность. В случае, когда модель не проходит эту проверку, изменить модель: например, провести логарифмирование отклика y (и пересчитать параметры модели уже для преобразованных значений y), расширить множество регрессоров, включенных в модель, в т.ч. и за счет незначимых.

5. Проверить качество построенной регрессионной модели с использованием метода кросс-валидации: сделать вывод о степени изменения значения коэффициента детерминации и показателя качества прогноза. Если такие изменения существенны, то модель нельзя считать качественной, и необходимо изменение ее структуры, либо анализ данных на предмет наличия выбросов в выборке.