Софийски университет „Св. Климент Охридски“ – гр. София Факултет по математика и информатика. Спец. Информационни системи – 3ти курс

КУРСОВ ПРОЕКТ

Статистика и емпирични методи

Практикум

(Задължителна дисциплина)

Регресионен анализ

(Линейна регресия)

Изготвил:

Петя Котова,

ФН. **71866**, курс 3

Преподавател:

Доц. д-р Деян Палежев

Проверил:

Методий Кандиларов

СЪДЪРЖАНИЕ:

1. Същност на анализа
2. Тема
3. Данни
4. Анализ на данните
   4.1.Едномерен анализ на всяка променлива
      4.1.1. Засичане на outlier-и
      4.1.2. Определяне на локацията на разпределението – median, mean, квартили
      4.1.3. Определяне на разсейването на разпределението –sd, range, IQR
      4.1.4. Графики
      4.1.5. Тестове за вида на разпределението
   4.2.Многомерен анализ
      4.2.1. Категорийна vs числова – One-way ANOVA и техните непараметрични еквиваленти.
      4.2.2. Числова vs числова – корелационен анализ, линейна регресия/ковариационен анализ.
      4.2.3. Графики
5. Заключение
6. Източници

1. СЪЩНОСТ НА АНАЛИЗА

Регресионния анализ се основава на въпроси дали съществува **функционална зависимост между две зависими случайни величини** и ако да – да се намери функция, която да я описва достатъчно точно. Класически пример е търсенето на зависимост между ръста и теглото на човек. Регресионният анализ не дава отговор на въпроса какви са причините. Той показва взаимните отношения между променливите, които в контекста на разглежданата задача могат да бъдат интерпретирани като причинно-следствени. Предназначен е за решаване на общи задачи – относно вида на зависимостта, определяне функцията на тази зависимост, количествено определяне параметрите на избраната функция. Променливите, чиито вариации искаме да обясним или предскажем, се наричат **зависими** – това е следствието. Целите на регресионния анализ са да определи как и в каква степен зависимата променлива варира или се променя като функция от изменения на независимата променлива, която е причината.

Линейният регресионен анализ се основава на шест основни предположения:

- Зависимите и независими променливи показват линейна връзка между наклона и прехващането.
- Независимата променлива не е случайна.
- Стойността на остатъка (грешка) е нула.
- Стойността на остатъка (грешка) е постоянна във всички наблюдения.
- Стойността на остатъка (грешка) не е свързана във всички наблюдения.
- Останалите стойности (грешка) следват нормалното разпределение.

Линейна регресия

Единичната (обикновена) линейна регресия описва линейната зависимост между независимата променлива x и зависимата променлива y (y = f(x)).Когато зависимостта между двете променливи (резултативната Y и факторната X) е линейна по форма, то точките са разположени около въображаема права линия (възходяща или низходяща). В този случай търсим уравнението на правата, която минава "най-близо" до точките от корелационното поле, т.е. най-добре отразява зависимостта между двете променливи. Критерий за "най-близо" – сборът от квадратите на разликите между емпиричните стойности *y* и техните оценки $\hat{y}$, които са ординатите на съответните точки от правата, да има минимум, т.е. $\sum (y-\hat{y})^2 = \text{minimum}$. Търсеното уравнение $\hat{y} = a + bx$ се нарича **регресионно уравнение (линеен регресионен модел).**

За намирането на неизвестните коефициенти *a* и *b* се прилага **методът на най-малките квадрати**, при което се стига до системата

$$\begin{cases} \sum y = Na + b\sum x \\ \sum xy = a\sum x + b\sum x^2 \end{cases}$$

и се получава решението $b = \frac{\sum xy - N\bar{x}\bar{y}}{\sum x^2 - N\bar{x}^2}$, $a = \bar{y} - b\bar{x}$.

След определянето на коефициентите *a* и *b* се получава **регресионният модел** $\hat{y} = a + bx$. Коефициентът *b* се нарича **регресионен коефициент** – той показва с колко единици се изменя зависимата променлива при изменение на факторната променлива с единица. Чрез регресионното уравнение могат да се получат оценките $\hat{y}$ за всяка стойност на *x*:

$\hat{y}_1 = a + bx_1$, $\hat{y}_2 = a + bx_2$, и т.н.

Множествената линейна регресия дава възможност да анализираме влиянието на две или повече независими променливи върху една зависима променлива.

Функцията изразяваща връзката между x_i и y:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots \beta_p x_p + \varepsilon$$

2. ТЕМА

Настоящият курсов проект има за цел да изследва разходите на новооткрити компании по такъв начин, че да успеем да определим каква ще бъде тяхната печалба.

Извадката, използвана в изследването, съдържа в себе си 50 наблюдения, репрезентативни изцяло за целите на този анализ. Семпълът, обект на настоящия проект, е извлечен от тук. Върху посочените данни е приложена множествена линейна регресия с очакван резултат свързан с определяне печалбата на компанията. За целите на изследването е използван статистически софтуер " R Studio", като биват показани последователно стъпките до достигане на финален модел, обясняващ най-добре таргетиранта променлива, а именно – **печалбата на съответната компания.**

3. ДАННИ

Пряк достъп до таблицата с данните ТУК:

- **RDSpend** - Разходите за научноизследователска и развойна дейност (НИРД) са пряко свързани с изследванията и развитието на фирмени стоки или услуги и всяка интелектуална собственост, генерирана в процеса.
- **Administration** - Административни разходи
- **Marketing Spend** - Маркетинговите разходи са общите разходи на организацията за маркетингови дейности.
- **State** - Щат
- **Profit** – Печалба

4. АНАЛИЗ НА ДАННИТЕ

- *Зареждане на всички нужни библиотеки и пакети*

```
library(Hmisc)
library(ggplot2)
library(corrplot)
library(caTools)
library(pR2)
library(car)
library(ROCR)
library(predictior
library(tseries)
library(fastDummie
library(readxl)
library(pscl)
library(stats)
```

```
install.packages("Hmisc")
install.packages("ggplot2")
install.packages("corrplot")
install.packages('caTools')
install.packages("pR2")
install.packages("car")
install.packages("ROCR")
install.packages('prediction')
install.packages("tseries")
install.packages('fastDummies')
```

- *Зареждане на основната таблица, съдържаща всички данни*

```
Data <- read_excel("C:/Users/Petya Kotova/Downloads/50_Startups.xlsx")
```

- *Трансформиране на данните в числови*

```
Data$RDSpend <- as.numeric(Data$RDSpend)
Data$Administration <- as.numeric(Data$Administration)
Data$MarketingSpend <- as.numeric(Data$MarketingSpend)
Data$Profit <- as.numeric(Data$Profit)
```

4.1. ЕДНОМЕРЕН АНАЛИЗ НА ВСЯКА ПРОМЕНЛИВА

4.1.1 ЗАСИЧАНЕ НА OUTLIER-и

```
lower_bound1 <- quantile(Data$RDSpend, 0.025)
lower_bound1

lower_bound2 <- quantile(Data$Administration, 0.025)
lower_bound2

lower_bound3 <- quantile(Data$MarketingSpend, 0.025)
lower_bound3
```

lower_bound1	lower_bound2	lower_bound3
2.5%	2.5%	2.5%
121.9613	54939.23	0

```
upper_bound1 <- quantile(Data$RDSpend, 0.975)
upper_bound1

upper_bound2 <- quantile(Data$Administration, 0.975)
upper_bound2

upper_bound3 <- quantile(Data$MarketingSpend, 0.975)
upper_bound3
```

upper_bound1	upper_bound2	upper_bound3
97.5%	97.5%	97.5%
160537.6	157436	435806.6

```
outlier_ind1 <- which(Data$RDSpend < lower_bound1 | Data$RDSpend > upper_bound1)
outlier_ind1
```

Изход: [1] 1 2 48 50

```
outlier_ind2 <- which(Data$Administration < lower_bound2 | Data$Administration > upper_bound2)
outlier_ind2
```

Изход: [1] 29 35 38 49

```
outlier_ind3 <- which(Data$MarketingSpend < lower_bound2 | Data$MarketingSpend > upper_bound2)
outlier_ind3
```

Изход: [1] 1 2 3 4 5 6 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 28 33 34 35 36 37 38 39 40 41 42 44 45 46 47 48 49 50

Outlier е наблюдение, което се намира на необичайно разстояние от други стойности в от популацията.В анализа използваме 97,5 percentile и 2,5 percentile, за да определим наблюденията – над и под тази стойност, като необичайни (outlier) за семпъла.

4.1.2 ОПРЕДЕЛЯНЕ НА ЛОКАЦИЯ НА РАЗПРЕДЕЛЕНИЕТО median, mean, квартили

```
summary(Data)
```

```
    RDSpend       Administration   MarketingSpend      State                Profit
Min.   :     0   Min.   : 51283   Min.   :     0   Length:50          Min.   : 14681
1st Qu.: 39936   1st Qu.:103731   1st Qu.:129300   Class :character   1st Qu.: 90139
Median : 73051   Median :122700   Median :212716   Mode  :character   Median :107978
Mean   : 73722   Mean   :121345   Mean   :211025                      Mean   :112013
3rd Qu.:101603   3rd Qu.:144842   3rd Qu.:299469                      3rd Qu.:139766
Max.   :165349   Max.   :182646   Max.   :471784                      Max.   :192262
```

	0%	25%	50%	75%	100%
quantile(Data$RDSpend)	0.00	39936.37	73051.08	101602.80	165349.20
quantile(Data$Administration)	51283.14	103730.88	122699.79	144842.18	182645.56
quantile(Data$MarketingSpend)	0.0	129300.1	212716.2	299469.1	471784.1

4.1.3 ОПРЕДЕЛЯНЕ НА РАЗСЕЙТВАНЕТО НА РАЗПРЕДЕЛЕНИЕТО sd, range, IQR

	RDSpend		
	КОД		Резултат
SD	**STD1<-sd(Data$RDSpend)**		45902.26
Range	**range1 <-range(Data$RDSpend)**	0.0	165349.2
IQR	**IQR1 <- IQR(Data$RDSpend)**		61666.43

	Administration		
	КОД		Резултат
SD	**STD2 <-sd(Data$Administration)**		28017.8
Range	**range2 <-range(Data$Administration)**	51283.14	182645.56
IQR	**IQR2 <- IQR(Data$Administration)**		41111.3

	Marketing Spending		
	КОД		Резултат
SD	**STD3 <- sd(Data$MarketingSpend)**		122290.3
Range	**range3 <-range(Data$MarketingSpend)**	0.0	471784.1
IQR	**IQR3 <- IQR(Data$MarketingSpend)**		170169

Median е средната стойност на група числа, подредени по големина. Тя е числото, което е точно в средата, така че 50% от класираните числа са над и 50% - под нея, още се дефинира като втори квартил.

Mean е средната(аритметична) стойност.

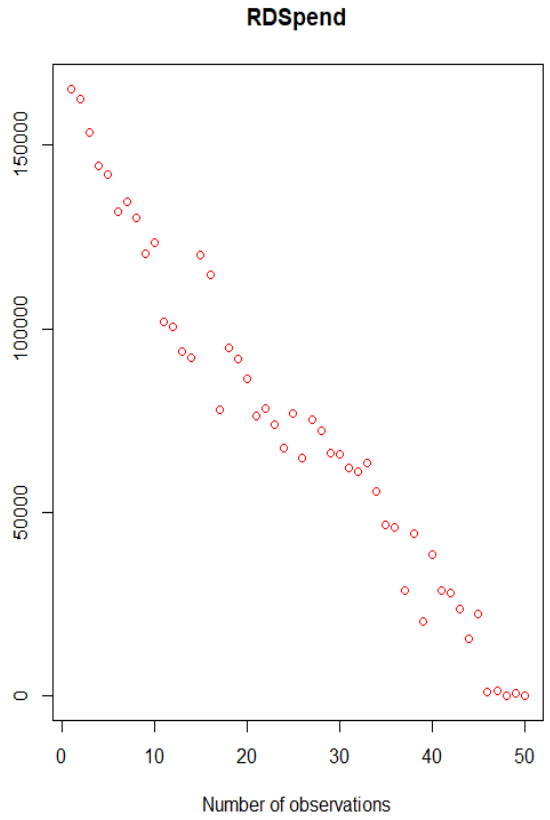
STD е мярка за размера на вариацията или дисперсията. Ниското стандартно отклонение показва, че стойностите са склонни да бъдат близки до средната стойност (наричана още очакваната стойност), докато високо стандартно отклонение показва, че стойностите са разпределени в по-широк диапазон.

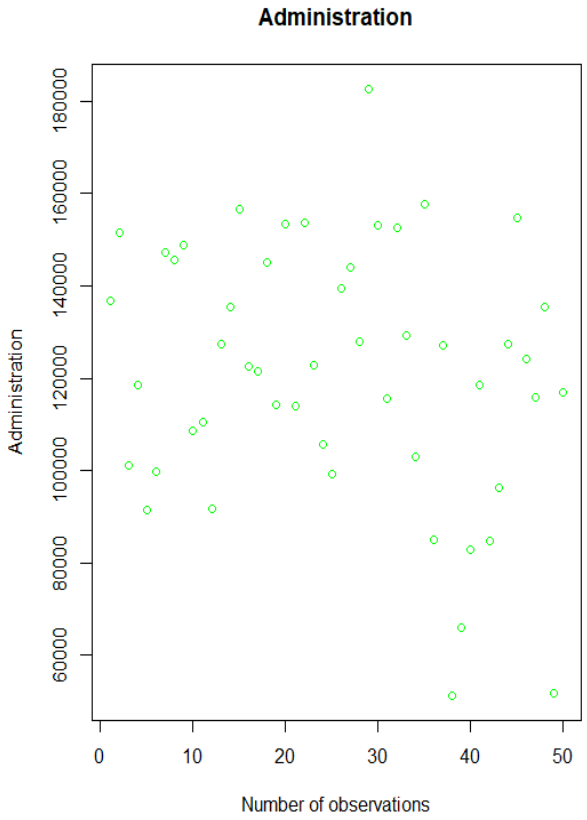
Range показва най-ниската и най-високата стойност в разпределението.

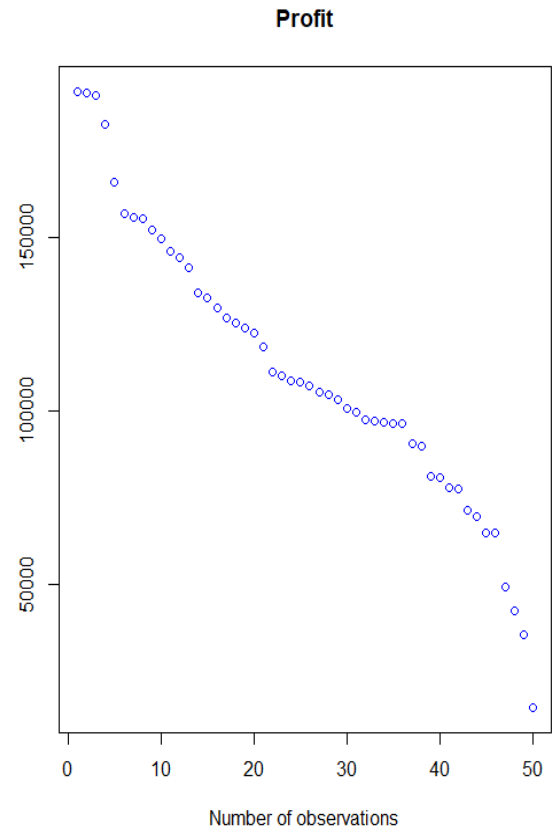
IQR е средните 50% от набора от данни. Това е диапазонът от стойности между третия квартил и първия квартил (Q3 - Q1).

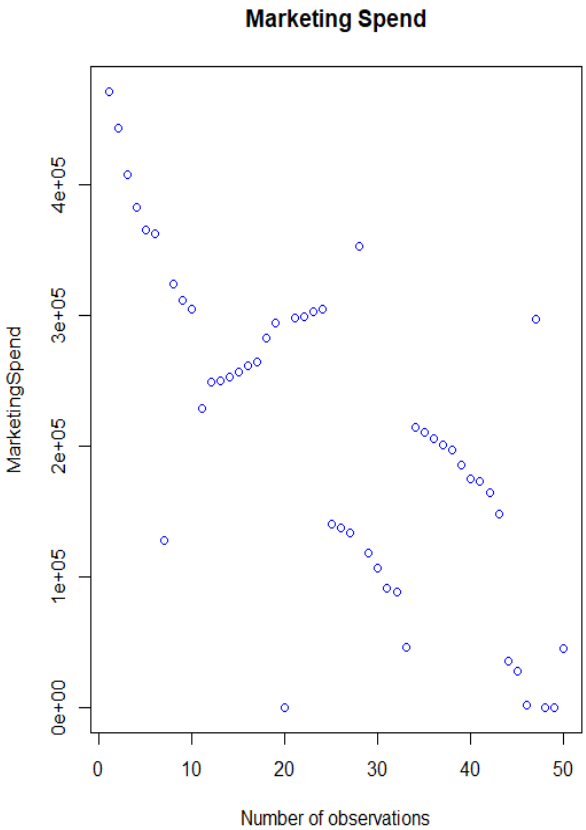
4.1.4. ГРАФИКИ

Разпределението на наблюденията на зависимата и независимите променливи, предмет на анализа е представено на на графиките.

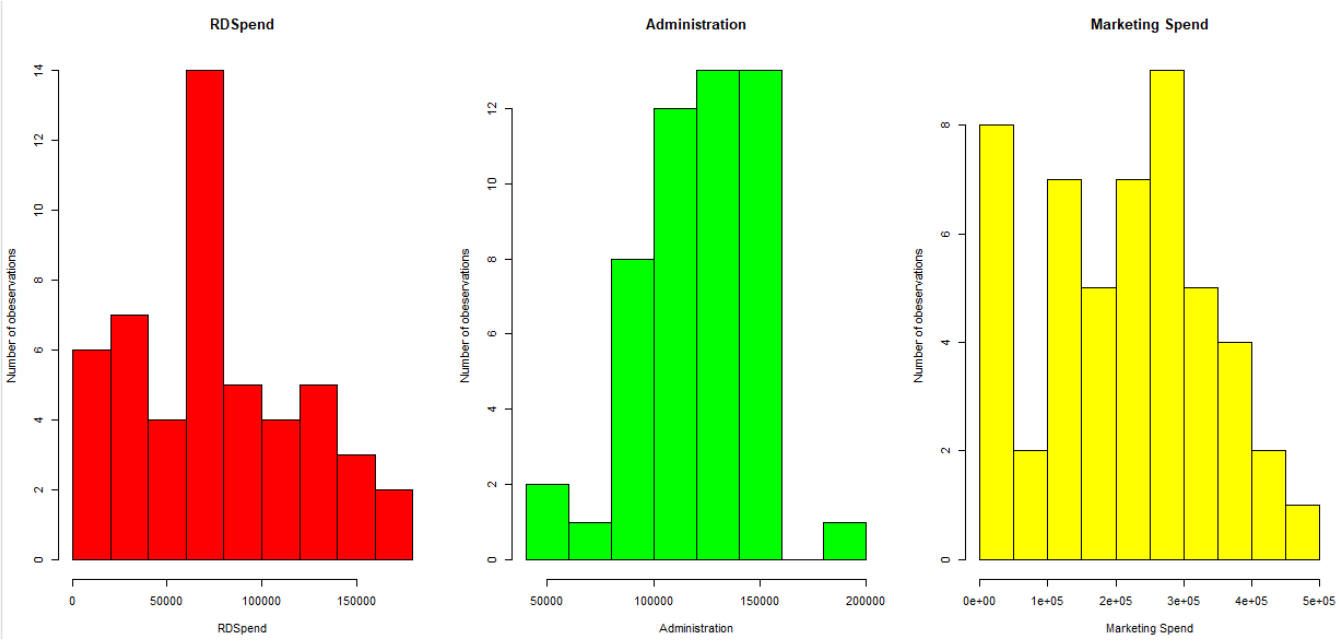






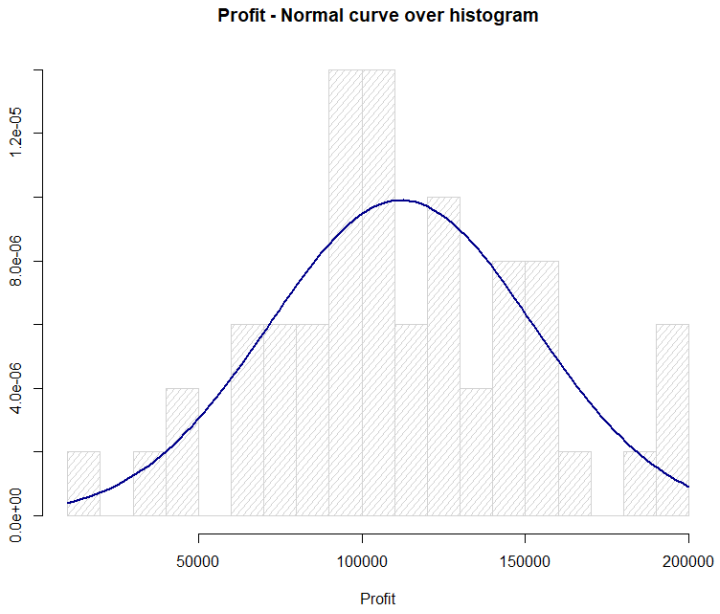


Хистограми на независимите числови променливи



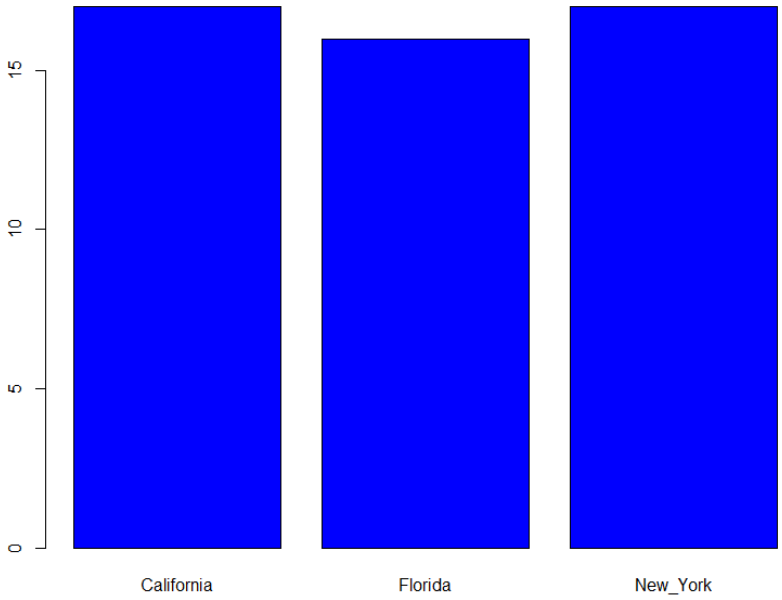
Хистограма на зависимата променлива:

```
m3<-mean(Data$Profit)
STD3 <-sd(Data$Profit)

hist(Data$Profit, density=20, breaks=20, prob=TRUE,
     xlab="Profit",
     main="Profit - Normal curve over histogram")
curve(dnorm(x, mean=m3, sd=STD3),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```



Построяваме хистограма и за зависимата променлива. Разпределението на наблюденията заедно с изобразяването на нормалното разпределение показва, че визуално може да определим, че зависимата променлива следва нормално разпределение.

Бар плот за категорийната променлива „State“



- ## Кодиране на категорийните данни.

Използването на категорийна променлива в линейна регресия изисква нейното кодиране. В частност методът, който е използван е чрез dummy variables.

```
Data <- dummy_cols(Data, select_columns = 'State')
```

4.1.5 ТЕСТОВЕ ЗА ВИДА НА РАЗПРЕДЕЛЕНИЕТО

SHAPIRO TEST			
Променлива	Код	W	P
RDSpend	shapiro.test(Data$RDSpend)	0.96734	0.1801
Administration	shapiro.test(Data$Administration)	0.97024	0.2366
Marketing Spend	shapiro.test(Data$MarketingSpend)	0.97437	0.3451

За изследване вида на разпределението използваме Shapiro-Wilk test(test for normality of the distribution).

За целта е важно да дефинираме нулевата и алтернативната хипотези.

H_0: “Данните са нормално разпределени “;

H_1: ” Данните не са нормално разпределени “.

Резултатите, които получаваме за независимите променливи, показват P value > 0,05 (ниво на съгласие), което означава, че нямаме основание да отхвърлим нулевата хипотеза H_o= “ Данните са нормално разпределени“.

4.2 МНОГОМЕРЕН АНАЛИЗ

4.2.1 КАТЕГОРИЙНА vs ЧИСЛОВА – One-way ANOVA И ТЕХНИТЕ НЕПАРАМЕТРИЧНИ ЕКВИВАЛЕНТИ

```
res.aov <- aov(RDSpend ~ State, data = Data)
summary(res.aov)
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
State         2 2.285e+09 1.143e+09   0.532  0.591
Residuals    47 1.010e+11 2.148e+09
```

```
res.aov <- aov(Administration ~ State, data = Data)
summary(res.aov)
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
State         2 9.634e+06   4816977   0.006  0.994
Residuals    47 3.846e+10 818196433
```

```
res.aov <- aov(MarketingSpend ~ State, data = Data)
summary(res.aov)
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
State         2 3.542e+10 1.771e+10   1.194  0.312
Residuals    47 6.974e+11 1.484e+10
```

```
res.aov <- aov(Profit ~ State, data = Data)
summary(res.aov)
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
State         2 1.901e+09 9.503e+08   0.575  0.567
Residuals    47 7.770e+10 1.653e+09
```

One-Way ANOVA ("analysis of variance") сравнява средните на две или повече независими групи, за да детерминира дали разликата в средните на популацията е статистически значима.

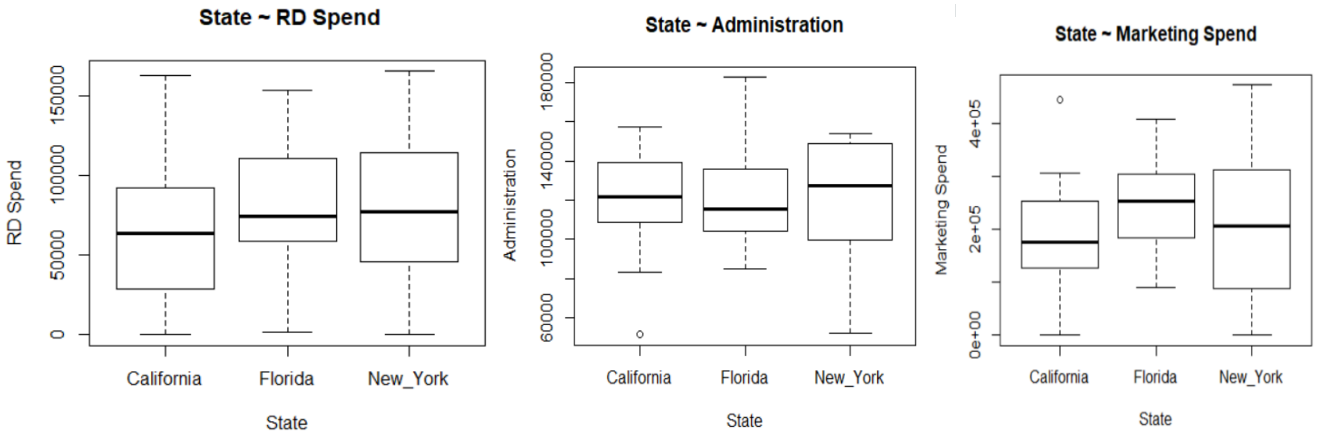
За целта дефинираме нулевата и алтернативната хипотеза:

H_0: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$

H_1: Поне едно μ_i е различно,

където μ_i средната на популацията в $i^{\text{-тата}}$ група ($i = 1, 2, ..., k$).
Резултатите, които получаваме, показват, че при ниво на съгласие 0,05 нямаме основание да отхвърлим нулевата хипотеза за всяка една от независимите променливи.

На графиките е представен box plot показващ разпределението на стойностите между числовите и категорийната променлива.

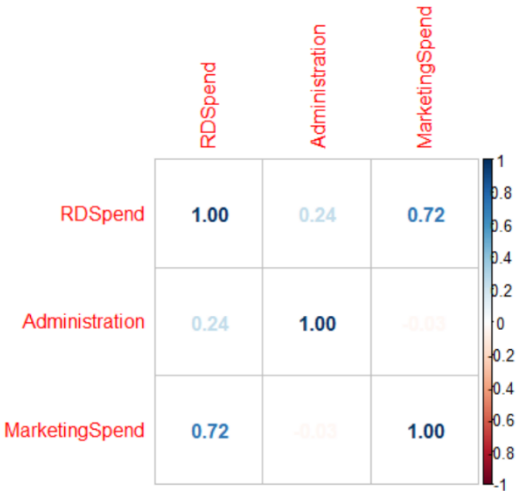


4.2.2 ЧИСЛОВА vs ЧИСЛОВА – КОРЕЛАЦИОНЕН АНАЛИЗ, ЛИНЕЙНА РЕГРЕСИЯ/КОВАРИЦИОНЕН АНАЛИЗ

- Създаване на матрица, която показва дали има корелация между променливите

```
rcorr(as.matrix(subset(Data, select=-c(Profit, State)), type=c("spearman")))
data_cor = subset(Data,select=c(RDSpend,Administration,MarketingSpend))
M = cor(data_cor[sapply(data_cor, function(x) !is.factor(x))])
corrplot(M, method="number")
```

```
                RDSpend Administration MarketingSpend
RDSpend            1.00           0.24           0.72
Administration     0.24           1.00          -0.03
MarketingSpend     0.72          -0.03           1.00

n= 50


P
                RDSpend Administration MarketingSpend
RDSpend                 0.0905         0.0000
Administration  0.0905                 0.8246
MarketingSpend  0.0000  0.8246
```



Корелационният анализ(силата на връзката между две променливи) показва, че се наблюдават значителна корелация между независимите променливи – RD Spend и Marketing Spend, което ще бъде взето предвид при построяването на модела.

- Създаване на матрица, която показва ковариацията на променливите

```
                 RDSpend Administration MarketingSpend
RDSpend       2107017150      311173891     4065495345
Administration 311173891      784997271     -110169009
MarketingSpend 4065495345    -110169009    14954920097
```

Ковариацията измерва общата вариация на две случайни величини от очакваните им стойности. Използвайки ковариация, можем само да преценим посоката на връзката (дали променливите са склонни да се движат в тандем или да показват обратна връзка). Това обаче не показва силата на връзката, нито зависимостта между променливите.

Разделяме данните от таблицата на **training set**, който ще бъде използван за моделиране и **test set**, върху който ще бъде приложен вече изградения модел, за да оценим как работи **регресионния модел** върху нови наблюдения, репрезентативни, но различни от **training set-a.**

```
set.seed(123)

#split the sample into training and test sets

sample = sample.split(Data$Profit, SplitRatio = .8)
train = subset(Data, sample == TRUE)
test  = subset(Data, sample == FALSE)
```

- Използваме **set.seed(123),** за да получаваме винаги еднакви части за **training set** и **test set,** като отношението на разделение е 80% за **training set** и 20% за **test set**.

- Скалиране на независимите променливи

Скалирането на независимите променливи е много често срещана практика при използването на multivariate analysis techniques.

Много техники приемат, че величината на измерването е пропорционална на неговата важност и че нивото на шума е сходно при всички променливи. Когато променливите имат значително различни мащаби, величината на стойностите не е непременно пропорционална на съдържанието на информацията. По същия начин, мащабът също е проблем, когато някои променливи съдържат повече шум от други променливи.

За да приложим scaling procedure върху нашите данни, използваме съответната функция в R studio.

```
training_set = subset(training_set, select = -c (State, State_New_York))
test_set = subset(test_set, select = -c(State, State_New_York))

training_set_scaled <- scale((subset(training_set, select=-c(Profit))))
test_set_scaled = scale((subset(test_set, select=-c(Profit))))

Profit1 <- subset(training_set, select = c (Profit))
Profit2 <- subset(test_set, select = c (Profit))

Training_set1=cbind(training_set_scaled, Profit1)

Test_set1=cbind(test_set_scaled, Profit2)
```

- Създаване на линеен регресионен модел

Първият модел, който ще бъде създаден, ще включва всички променливи налични в нашия семпъл.

```
# run a regression which includes all the vars
regressor =lm(formula = Profit ~., data=Training_set1)
#take a look at the results
summary(regressor)

Call:
lm(formula = Profit ~ ., data = Training_set1)

Residuals:
   Min     1Q Median     3Q    Max
-31659  -4371   -307   5362  18340

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      109736.8     1463.0  75.005   <2e-16 ***
RDSpend           39118.9     2267.5  17.252   <2e-16 ***
Administration      388.1     1573.1   0.247    0.807
MarketingSpend     2410.7     2254.3   1.069    0.292
State_California    -52.5     1700.0  -0.031    0.976
State_Florida       453.4     1714.5   0.264    0.793
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9253 on 34 degrees of freedom
Multiple R-squared:  0.9575,    Adjusted R-squared:  0.9512
F-statistic: 153.1 on 5 and 34 DF,  p-value: < 2.2e-16
```

Резултатът от регресията показва, че единствената променлива, която влиза в регресията е RDSpend (при ниво на съгласие 0,05), за останалите независими променливи наблюдаваме, че те не са статистически значими за модела. Затова преминаваме към втори challenger модел, които включва само тази променлива.

- Създаване на втори линеен регресионен модел

```
regressor2 = lm(formula = Profit ~ RDSpend, data=Training_set1)
summary(regressor2)
```

```
Call:
lm(formula = Profit ~ RDSpend, data = Training_set1)

Residuals:
   Min     1Q Median     3Q    Max
-32118  -4844   -287   6377  17252

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   109737       1413   77.69   <2e-16 ***
RDSpend        40952       1431   28.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8934 on 38 degrees of freedom
Multiple R-squared:  0.9557,    Adjusted R-squared:  0.9545
F-statistic: 819.5 on 1 and 38 DF,  p-value: < 2.2e-16
```

Резултатите от регресията показват, че променливата е статистически значима(при ниво на съгласие 0,05) и достатъчна, за да бъде построен моделът.

Определяме втория модел като финален за нашето изследване с независима променлива влизаща в модела „ RD Spend” и R Squared, Adj R Squared равни на 95.5 %.

Моделът определен като финален бива наложен върху тестовия ни семпъл, като виждаме много близки предиктивни резултатите спрямо наблюдаваните такива.

```
y_pred = predict(regressor2, newdata = Test_set1)
y_pred
```

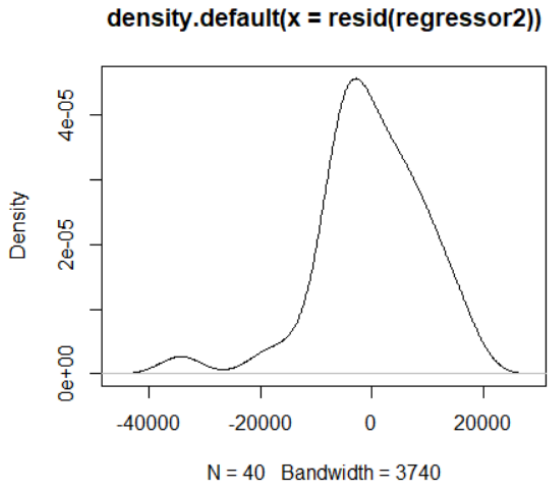
```
       1        2        3        4        5        6        7        8        9       10
178126.62 147204.11 146640.83 107229.75 122873.87 120140.58  86063.48  60845.17  75325.80  52917.70
```

- Допълнителни методи за оценка на модела:

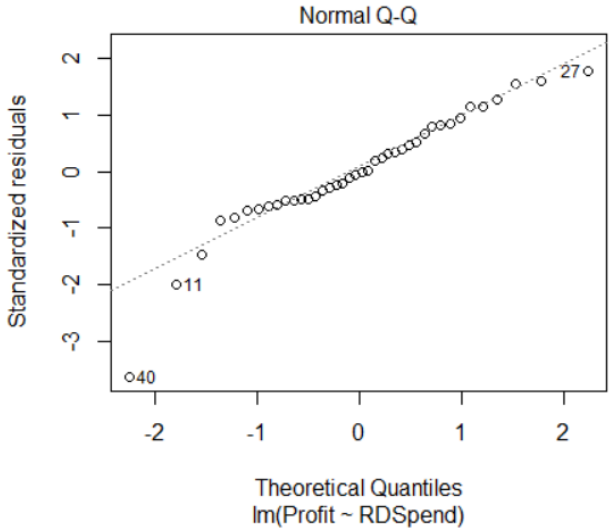
```
#r2ML  Maximum likelihood pseudo r-squared (Cox & Snell)
#r2CU   Cragg and Uhler's or NRDSpendlkerke's pseudo r-squared.

#averRDSpend prediction, looking at r2 evaluations
pR2(regressor2)
```

```
fitting null model for pseudo-r2
         llh      llhNull           G2     McFadden         r2ML         r2CU
-419.6360709 -481.9636520  124.6551622    0.1293201    0.9556827    0.9556827
```

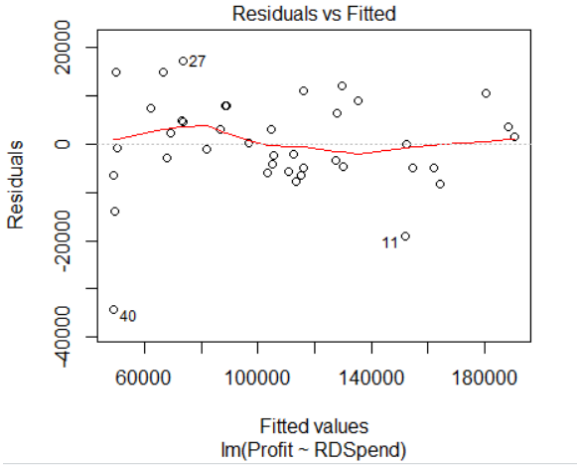
- Графики



Residual density plot открива поведението на остатъците. Функцията връща графика с прогнозна плътност на остатъците. Стойностите им се показват по оста x. За някои модели очакваната форма на плътност може да бъде получена от предположенията за модела. Например, остатъците от прости линейни модели трябва да бъдат нормално разпределени. Въпреки това, дори ако моделът няма предположение за разпределението на остатъците, графика на остатъчната плътност може да бъде информативен източник. Ако повечето от остатъците не са концентрирани около нулата, вероятно е прогнозите на модела да са предубедени.



Графиката Q-Q или квантилно-квантилната графика е графичен инструмент, който ни помага да оценим дали набор от данни е дошъл от някакво теоретично разпределение - като нормално. Създадена е чрез нанасяне на два квантили един срещу друг. Ако и двата набора квантили идват от едно и също разпределение, трябва да видим точките, образуващи права, която е приблизително права.



На графиката е показана прогнозираната стойност (червената линия), както и разпределението на остатъците (точките). Графиката се използва за откриване на нелинейност, неравномерни вариации на грешки и outliers.

5. ЗАКЛЮЧЕНИЕ:

Печалбата на стартиращи компании бе определена като зависима променлива, докато RD, административни, маркетинг разходи и щата са определени като независими променливи. Настоящия анализ цели създаването на линейна регресия, като да предскаже зависимата променлива чрез независимите променливи. Като първия създаден модел включва всички променливи налични в извадката, за да се провери поведението на независимите променливи и тяхната значимост, без да се имат предвид резултатите от корелационната зависимост. Вторият– финален модел, включва статистически значимата променлива (съответно на база на p-value стойността, от резултатите от регресията) „RD Spend“.

Изграденият финален модел показва стойности на R Squared, Adj R Squared равни на 95.5 %.

6. ИЗТОЧНИЦИ:

https://expert-bg.org/machine-learning-kakvo-e-regresia/

http://kb.smetni.com/I_4_1_Kabaivanov_Ikonometria_za_finansisti.pdf

https://bg.pharoskc.com/12-what-is-covariance

http://wiki.eigenvector.com/index.php?title=Advanced_Preprocessing:_Variable_Scaling