VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

TAIKOMOSIOS STATISTIKOS

TIRIAMASIS DARBAS

„Startuolių verslo skyriaus išlaidų ir pelno sąryšis“

Atliko: Ugnius Alekna

Matematika ir matematikos taikymai

III kursas

Turinys

Įžanga

Šis tyrimo darbas siekia išanalizuoti, kaip skirtingos verslo skyriaus išlaidos yra susijusios su startuolių (naujų įmonių, verslo projektų) pelningumu. Tyrimo metu bus siekiama ištirti duomenų sąryšį ir sukurti daugialypės regresijos modelį, kuris galėtų tiksliai prognozuoti naujų startuolių pelną, suteikdamas vertingų žinių investuotojams ir suinteresuotiems asmenims, siekiantiems įvertinti šių startuolių finansines perspektyvas.

Tyrimo tikslas: Ištirti startuolių pelno ir MTEP (mokslinių tyrimų ir eksperimentinės plėtros) išlaidų, administravimo išlaidų ir rinkodaros išlaidų sąryšį. Sukurti daugialypės regresijos modelį, kuris galėtų tiksliai numatyti startuolio pelną pagal atitinkamas įmonės (veiklos) išlaidas.

Norint pasiekti tyrimo tikslą, iš internetinės duomenų mokslo svetainės Kaggle.com buvo surinkti ir išanalizuoti duomenys apie 50-ies startuolių pelną bei įvairias įmonių (veiklų) išlaidas. Duomenys buvo ištirti pasitelkiant įvairius statistinės analizės metodus. Remiantis ištirtais duomenimis buvo sudaromas optimalus daugialypės regresijos modelis, kuris buvo pritaikytas prognozuojant hipotetinių startuolių pelną.

1. Duomenų apdorojimas ir statistinė analizė

1.1 Duomenys

Duomenų rinkinį „50_Startups.csv“ sudaro finansinė 50-ies JAV startuolių informacija. Duomenų rinkinyje yra 4 kintamieji, kurie suteikia įžvalgas apie startuolių charakteristikas bei finansinius rezultatus. Toliau pateikiamas išsamus duomenų rinkinyje esančių kintamųjų aprašymas:

- `R.D.Spend` – MTEP (mokslinių tyrimų ir eksperimentinės plėtros) išlaidos. Šis kiekybinis intervalinis kintamasis parodo pinigų sumą, kurią kiekvienas startuolis skyrė moksliniams tyrimams bei plėtrai. Šis kintamasis atspindi įmonės investicijas į inovacijas ir projekto kūrimą.
- `Administration` – administravimo išlaidos. Šis kiekybinis intervalinis kintamasis nurodo kiekvieno startuolio administravimo išlaidas, t.y. išlaidas, apimančias atlyginimus, biuro nuomą, įrangos ar technologijų pirkimus bei kitą organizacinę veiklą.
- `Marketing.Spend` – rinkodaros išlaidos. Tai pinigų suma, skirta rinkodaros veiklai, tokiai kaip reklama ar prekės ženklo kūrimas.
- `Profit` – pelnas. Tai priklausomas kiekybinis intervalinis kintamasis, kuris nurodo kiekvieno startuolio pelną. Šį kintamąjį galima interpretuoti kaip startuolio finansinės sėkmės matą.

Duomenų rinkinio (lentelės) `dataset` pirmas šešias eilutes galime atspausdinti naudodami komandą `head()`

```
> head(dataset)
  R.D.Spend Administration Marketing.Spend   Profit
1  165349.2       136897.80        471784.1 192261.8
2  162597.7       151377.59        443898.5 191792.1
3  153441.5       101145.55        407934.5 191050.4
4  144372.4       118671.85        383199.6 182902.0
5  142107.3        91391.77        366168.4 166187.9
6  131876.9        99814.71        362861.4 156991.1
```

1.2 Duomenų empirinės charakteristikos

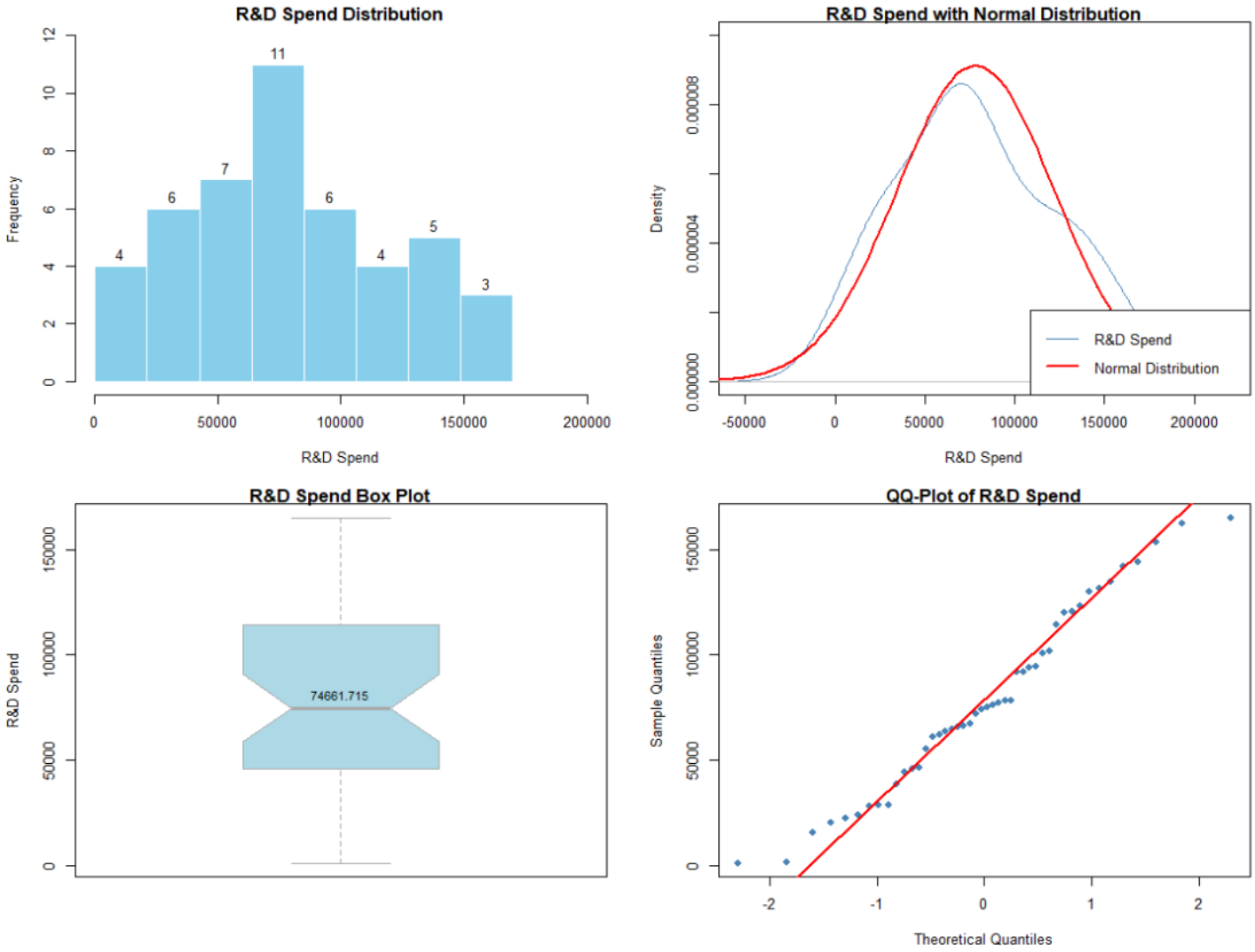
Norint analizuoti turimus duomenis, pirmiausia turime apskaičiuoti visų duomenų rinkinyje esančių kintamųjų empirines charakteristikas. Tai galime padaryti naudodami komandą `summary()`, kuri pateikia trumpą duomenų lentelės apžvalgą. Joje apskaičiuojamos įvairios kiekvieno kintamojo empirinės charakteristikos.

```
> summary(dataset)
   R.D.Spend      Administration   Marketing.Spend      Profit
 Min.   :   542   Min.   : 51283   Min.   :  1904   Min.   : 14681
 1st Qu.: 45528   1st Qu.:103731   1st Qu.:139269   1st Qu.: 90139
 Median : 74662   Median :122700   Median :229161   Median :107978
 Mean   : 76793   Mean   :121345   Mean   :224495   Mean   :112013
 3rd Qu.:105066   3rd Qu.:144842   3rd Qu.:301528   3rd Qu.:139766
 Max.   :165349   Max.   :182646   Max.   :471784   Max.   :192262
 NA's   :2                         NA's   :3
```

Iš trumpos duomenų apžvalgos matome, jog tarp kintamųjų `R.D.Spend` ir `Marketing.Spend` yra trūkstamų reikšmių. Tolimesnei analizei eilutes su trūkstamomis reikšmėmis apdorosime, pakeičiant jas atitinkamo kintamojo stulpelio vidurkio reikšme. Duomenis su apdorotomis trūkstamomis reikšmėmis išsaugome naujoje duomenų lentelėje `data`. Galime apžvelgti kiekvieno iš kintamųjų vidurkius bei standartinius nuokrypius. Taip pat galime pastebėti, jog kintamųjų `R.D.Spend` ir `Marketing.Spend` `Min.` reikšmės yra labai mažos, lyginant su vidurkiu, todėl tai – galimos išskirtys. Norint nustatyti, ar šie stebiniai tikrai yra išskirtys, reikalinga išsamesnė analizė.

1.3 Kintamųjų grafinis pavaizdavimas

1.3.1 MTEP išlaidų duomenų pavaizdavimas



`R.D.Spend` histograma ir teorinės tankio funkcijos grafikas parodo, jog duomenų pasiskirstymas yra panašus į varpo formos pasiskirstymą.

Duomenyse esantiems kintamiesiems apskaičiuosime asimetrijos koeficientą (`skewness`) ir ekscesą (`kurtosis`). Šios charakteristikos mums parodo, ar daug tiriamasis skirstinys skiriasi nuo normaliojo skirstinio su tokiu pačiu vidurkiu ir standartiniu nuokrypiu.

```
> skewness(data$R.D.Spend)
[1] 0.2068315
> kurtosis(data$R.D.Spend)
[1] 2.216607
```

Gautas `R.D.Spend` duomenų asimetrijos koeficientas lygus 0.2. Vadinasi, asimetrija yra dešinioji, tačiau ji nežymi, kadangi asimetrijos koeficientas artimas 0.
Gauta `R.D.Spend` duomenų eksceso reikšmė lygi 2.22. Tai reiškia, jog skirstinys yra smailesnis, negu normaliosios kreivės.

Asimetrijos bei eksceso charakteristikų palyginimui nubrėžiame duomenų teorinės tankio funkcijos ir normaliojo skirstinio grafikus.

Skirstinių normalumui tirti ir vizualizuoti taip pat galime panaudoti dėžinį grafiką ir kvantilių grafiką. Dėžinis grafikas naudojamas duomenų asimetriškumui (bet ne tankio formai) pavaizduoti. Linija, esanti žemiau vidurio, rodo duomenų medianą. Matome, jog `R.D.Spend` duomenų medianos linija yra ne per vidurį, tai paaiškina duomenų dešininį asimetriškumą. Figūrų apatinė ir viršutinė kraštinės rodo duomenų kvartilius, o brūkšniukai viršuje ir apačioje rodo trijų standartinių nuokrypių nuo vidurkio ribą. Taip pat iš dėžinio grafiko matome, jog duomenyse nėra vizualių išskirčių. Išskirčių nebuvimą detaliau patikrinsime truputį vėliau.

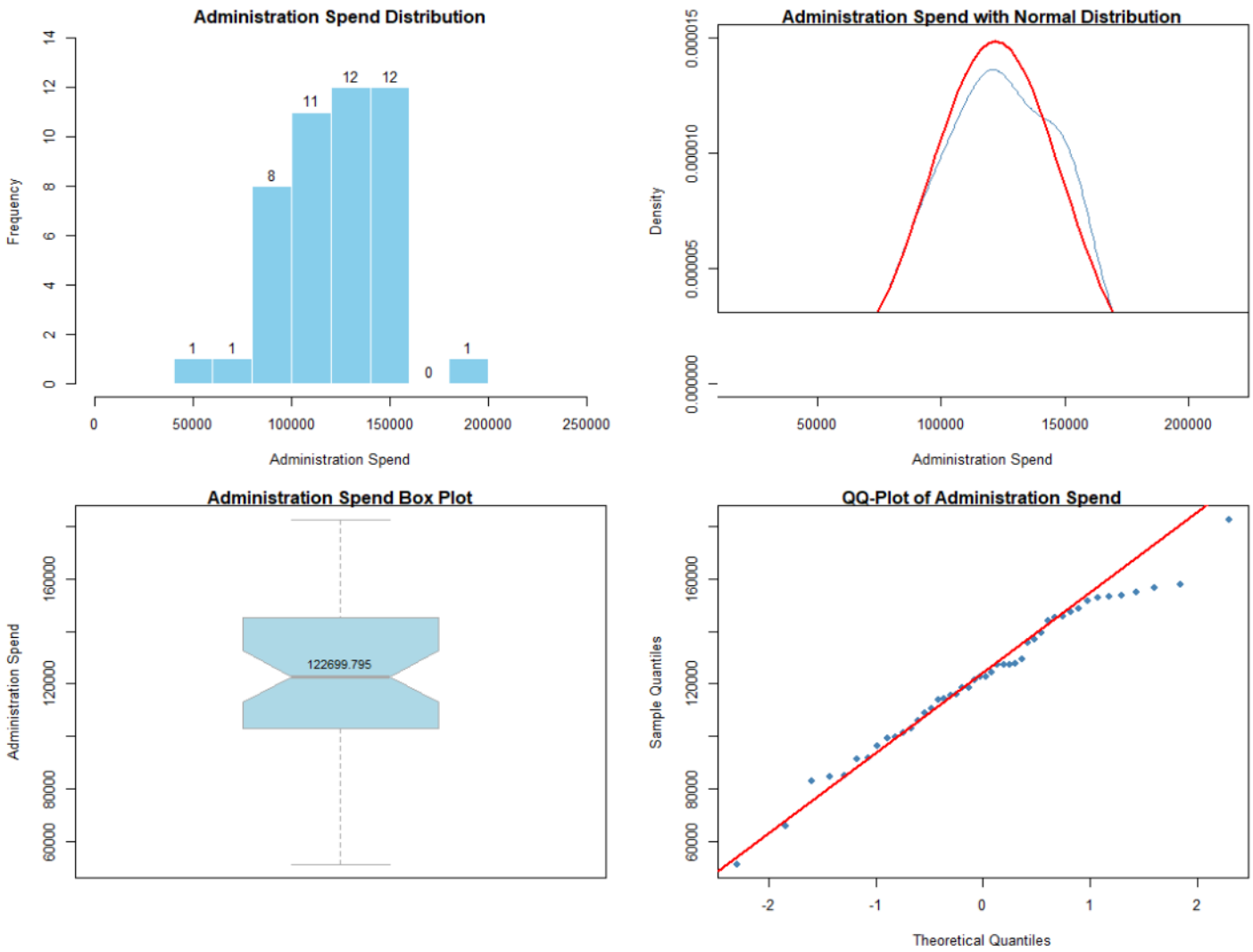
Iš kvantilių grafiko matome, jog `R.D.Spend` duomenų kvantiliai yra išsidėstę gan arti teorinių kvantilių tiesės, tačiau ne visi kvantiliai yra artimi tiesei. Kuo duomenys arčiau kraštų, tuo labiau nutolę nuo teorinių kvantilių tiesės. Panašu, jog duomenys yra artimi normaliesiems.

Patikrinti, ar tiriami duomenys yra iš normaliosios populiacijos galime naudodami Šapiro-Vilko testą. Šapiro-Vilko testo nulinė hipotezė teigia, jog duomenys yra iš normaliosios populiacijos. Vadinasi, kriterijaus `p-value` > 0.05 rodo, kad nulinę hipotezę galime priimti ir teigti, jog duomenys turi normalųjį skirstinį.

```
> shapiro.test(data$R.D.Spend)

        Shapiro-Wilk normality test

data:  data$R.D.Spend
W = 0.97178, p-value = 0.3221
```

Gauta Šapiro-Vilko kriterijaus `p-value` lygi 0.322 > 0.05. Vadinasi, galime teigti, jog duomenys yra iš normaliosios populiacijos.

1.3.2 Administravimo išlaidų duomenų pavaizdavimas



Administravimo išlaidų histograma ir teorinės tankio funkcijos grafikas parodo, jog duomenų pasiskirstymas yra panašus į varpo formos pasiskirstymą.

Kintamiesiems `Administration` apskaičiuosime asimetrijos koeficientą ir ekscesą.

```
> skewness(data$Administration)
[1] -0.3068269
> kurtosis(data$Administration)
[1] 2.957412
```

Gautas `Administration` duomenų asimetrijos koeficientas lygus -0.3. Vadinasi, asimetrija yra kairioji, tačiau ji nežymi, kadangi asimetrijos koeficientas artimas 0.
Gauta `Administration` duomenų eksceso reikšmė lygi 2.96. Tai reiškia, jog skirstinys yra smailesnis, negu normaliosios kreivės.

Tikslesniam palyginimui nubraižome duomenų teorinės tankio funkcijos ir normaliojo skirstinio grafikus.

Skirstinių normalumui tirti ir vizualizuoti taip pat galime panaudoti dėžinį grafiką ir kvantilių grafiką. Administravimo išlaidų duomenų medianos linija yra ganėtinai tiksliai viduryje, tai parodo, jog duomenys visgi yra simetriškai pasiskirstę.

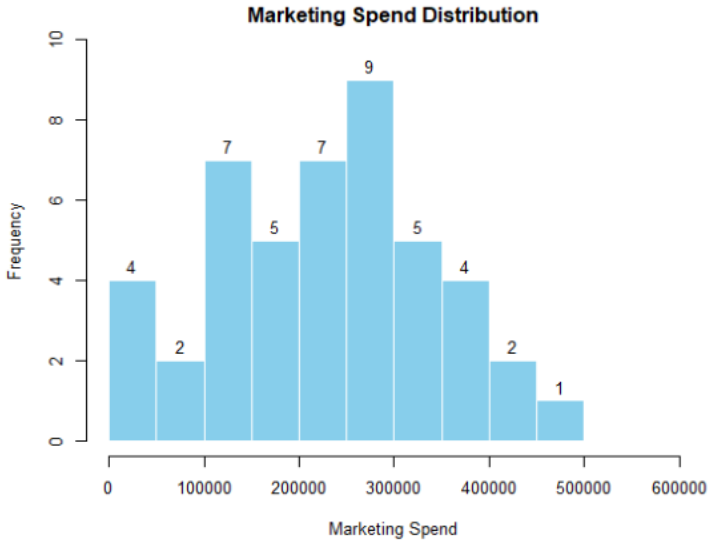
Matome, jog `Administration` duomenų kvantiliai yra išsidėstę gan arti teorinių kvantilių tiesės, tačiau ne visi kvantiliai yra artimi tiesei. Kuo duomenys arčiau kraštų, tuo labiau nutolę nuo teorinių kvantilių tiesės. Panašu, jog duomenys yra artimi normaliesiems.

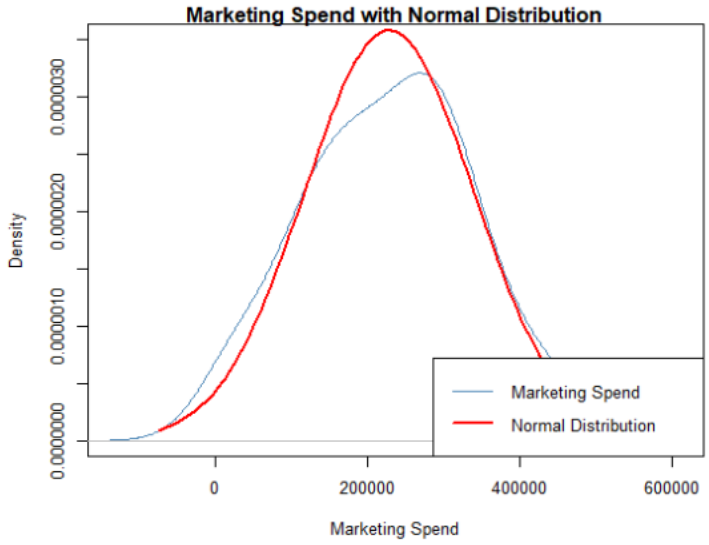
Patikrinti, ar tiriami duomenys yra iš normaliosios populiacijos galime naudodami Šapiro-Vilko testą.

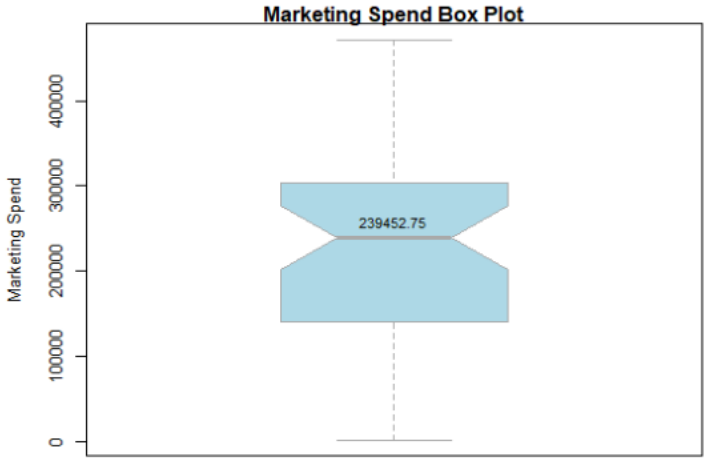
```
> shapiro.test(data$Administration)

        Shapiro-Wilk normality test

data:  data$Administration
W = 0.98253, p-value = 0.7106
```

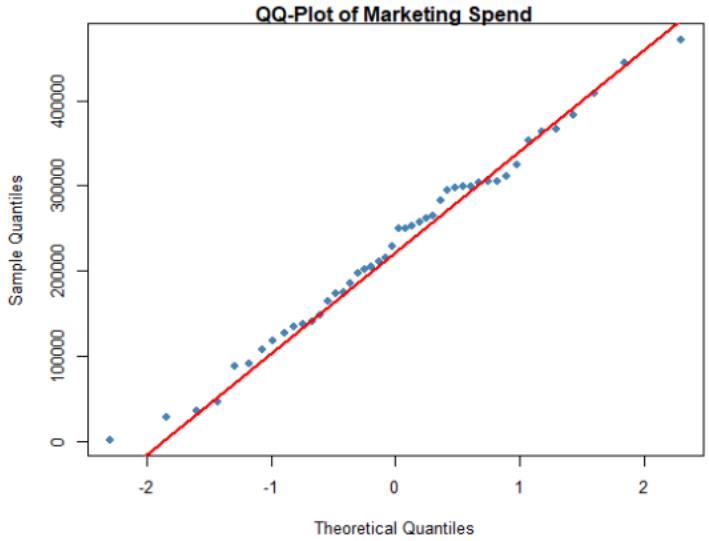
Gauta Šapiro-Vilko kriterijaus `p-value` lygi 0.711 > 0.05. Vadinasi, galime teigti, jog duomenys yra iš normaliosios populiacijos.

1.3.3 Rinkodaros išlaidų duomenų pavaizdavimas









`Marketing.Spend` histograma ir teorinės tankio funkcijos grafikas parodo, jog duomenų pasiskirstymas yra panašus į varpo formos pasiskirstymą.

Kintamajam `Marketing.Spend` apskaičiuosime asimetrijos koeficientą ir ekscesą.

```
> skewness(data$Marketing.Spend)
[1] -0.008942677
> kurtosis(data$Marketing.Spend)
[1] 2.477977
```

Gautas `Marketing.Spend` duomenų asimetrijos koeficientas lygus -0.009. Vadinasi, asimetrijos beveik nėra.
Gauta `Marketing.Spend` duomenų ekscesо reikšmė lygi 2.48. Tai reiškia, jog skirstinys yra smailesnis, negu normaliosios kreivės.

Skirstinių normalumui tirti ir vizualizuoti panaudojame dėžinį grafiką ir kvantilių grafiką.

`Marketing.Spend` duomenų medianos linija yra aukščiau vidurio, vadinasi, jog duomenų viršūnės asimetrija yra kairioji. Taip pat iš dėžinio grafiko matome, jog duomenyse nėra vizualių išskirčių. Išskirčių nebuvimą detaliau patikrinsime truputį vėliau.

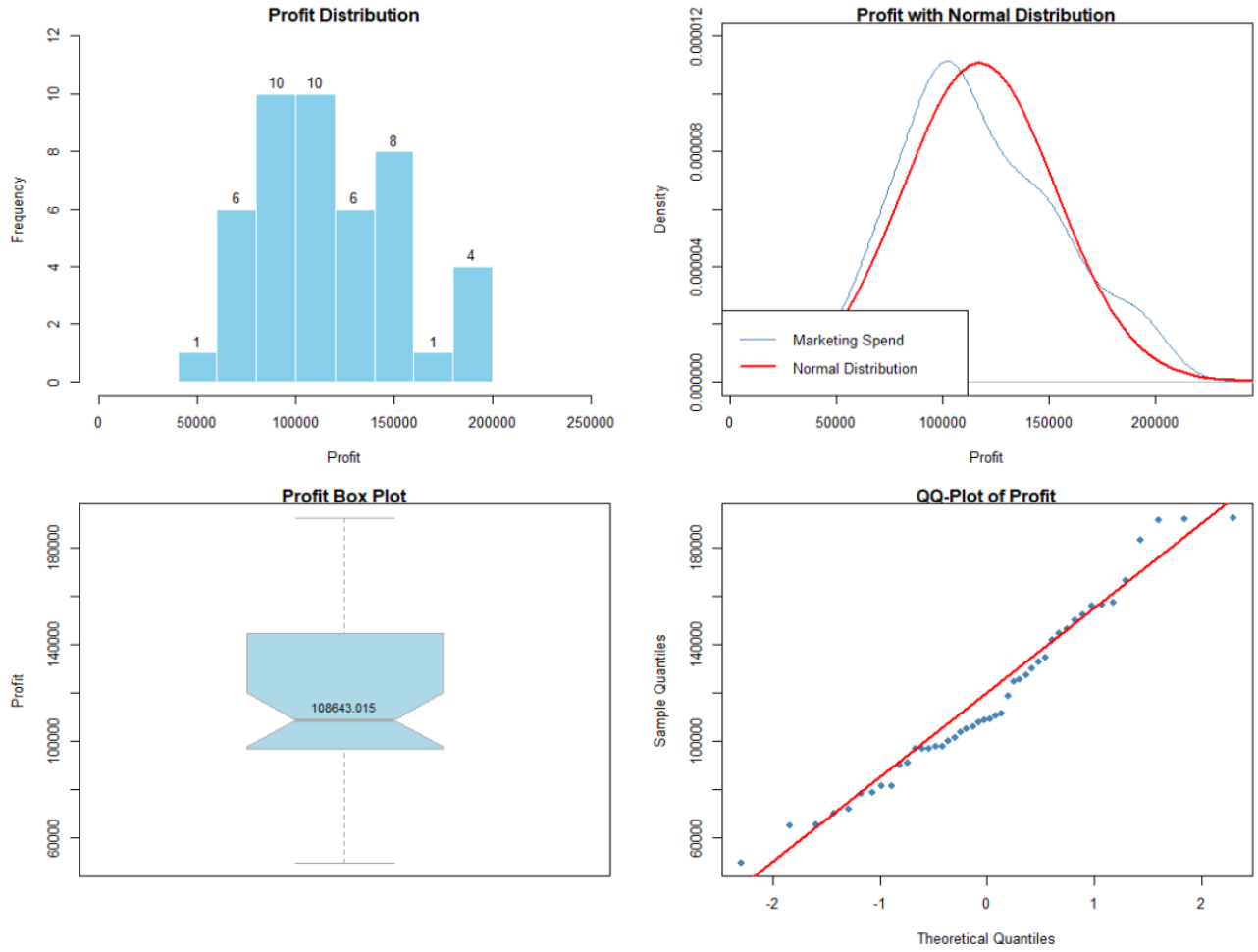
Matome, jog `Marketing.Spend` duomenų kvantiliai yra išsidėstę labai arti teorinių kvantilių tiesės. Keli kraštiniai duomenys yra šiek tiek nutolę nuo teorinių kvantilių tiesės. Duomenys yra artimi normaliesiems.

Patikrinti, ar tiriami duomenys yra iš normaliosios populiacijos galime naudodami Šapiro-Vilko testą.

```
> shapiro.test(data$Marketing.Spend)

        Shapiro-Wilk normality test

data:  data$Marketing.Spend
W = 0.98799, p-value = 0.9115
```

Gauta Šapiro-Vilko kriterijaus `p-value` lygi 0.912 > 0.05. Vadinasi, labai drąsiai galime teigti, jog duomenys yra iš normaliosios populiacijos.

1.3.4 Pelno duomenų pavaizdavimas



Pelno išlaidų histograma ir teorinės tankio funkcijos grafikas parodo, jog duomenų pasiskirstymas yra panašus į varpo formos pasiskirstymą.

Duomenyse esančiam kintamajam `Profit` apskaičiuojame asimetrijos koeficientą ir ekscesą.

```
> skewness(data$Profit)
[1] 0.4274668
> kurtosis(data$Profit)
[1] 2.502276
```

Gautas `Profit` duomenų asimetrijos koeficientas lygus 0.43.Vadinasi, jog duomenų asimetrija yra dešinioji.
Gauta `Profit` duomenų ekscesо reikšmė lygi 2.50. Tai reiškia, jog skirstinys yra smailesnis, negu normaliosios kreivės.

Skirstinių normalumui vizualizuoti panaudojame dėžinį grafiką ir kvantilių grafiką.

`Profit` duomenų medianos linija yra gerokai žemiau vidurio, vadinasi, jog duomenų asimetrija yra dešinioji.

Matome, jog Rinkodaros išlaidų duomenų kvantiliai yra išsidėstę ganėtinai arti teorinių kvantilių tiesės, tačiau yra ir šiek tiek nutolusių duomenų. Panašu, jog duomenys yra artimi normaliesiems, tačiau patikrinti, ar duomenys yra iš normaliosios populiacijos reikės Shapiro Vilko testo.

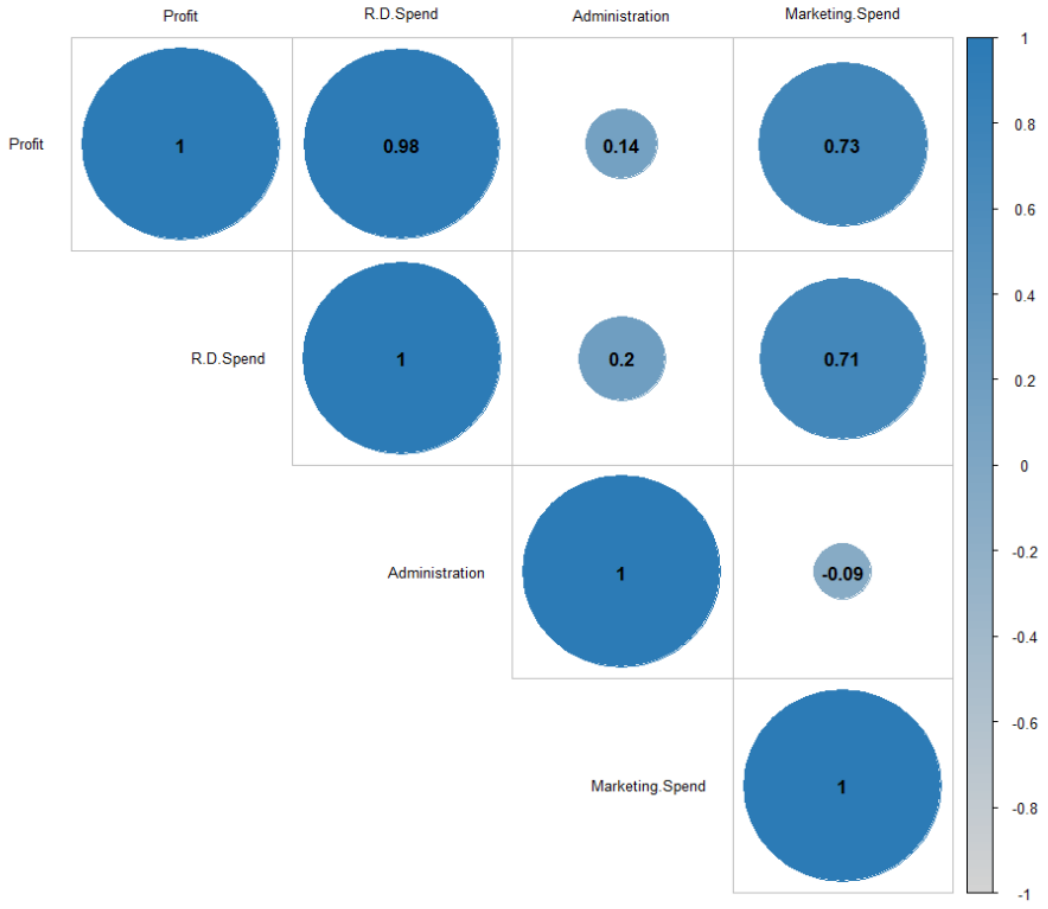
```
> shapiro.test(data$Profit)

        Shapiro-Wilk normality test

data:  data$Profit
W = 0.96436, p-value = 0.1693
```

Gauta Šapiro-Vilko kriterijaus `p-value` lygi 0.169 > 0.05. Vadinasi, galime teigti, jog duomenys yra iš normaliosios populiacijos.

1.4 Duomenų tarpusavio priklausomybė

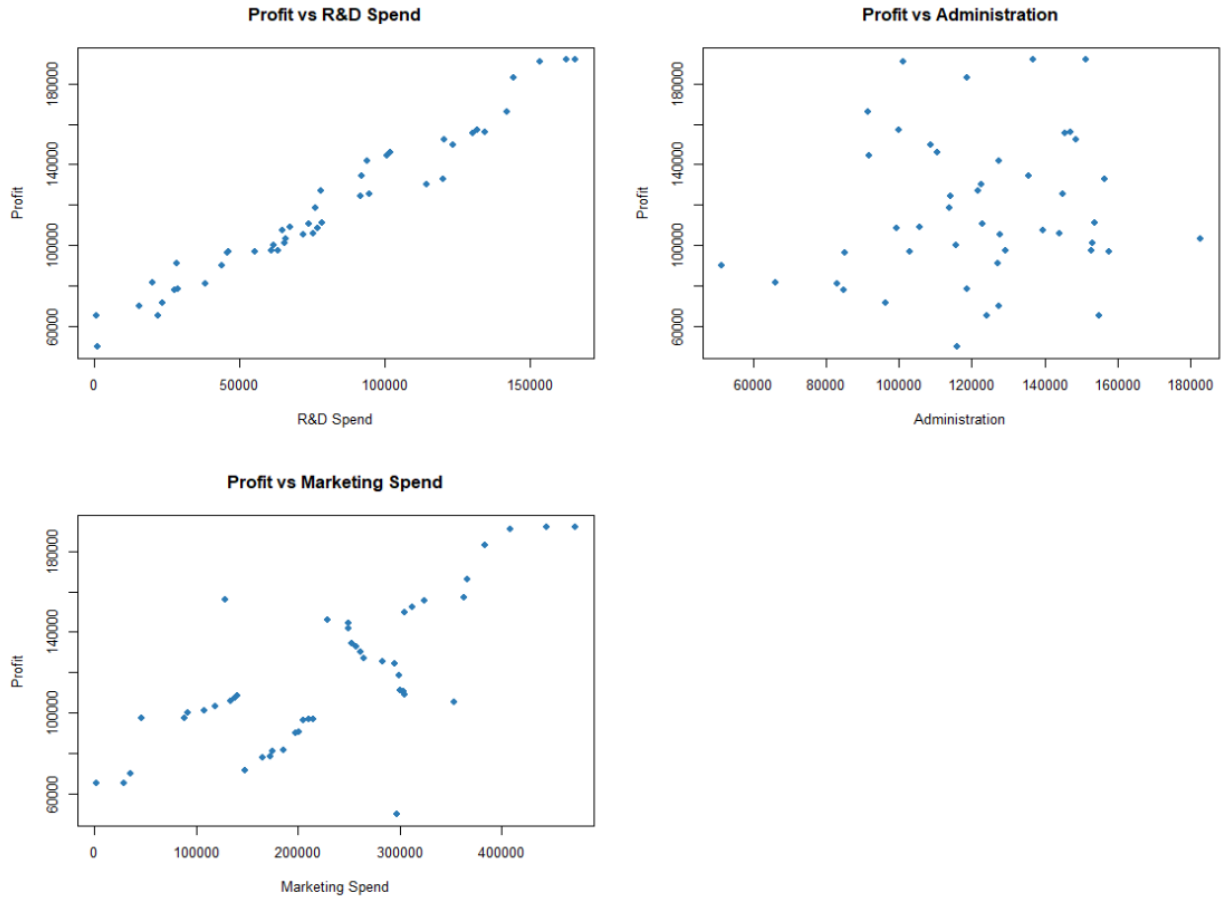
Norint ištirti tiesinę duomenų tarpusavio priklausomybę naudojama Pirsono koreliacijos koeficiento reikšmė. Koreliacijos koeficiento reikšmės yra iš intervalo [-1; 1]. Kuo koeficientas artimesnis 0, tuo tiesinis sąryšis tarp dviejų kintamųjų silpnesnis.

Norint pavaizduoti kiekvieno kintamojo tarpusavio koreliaciją patogu naudoti koreliacijos matricą.



Iš koreliacijos matricos matome, jog `Profit` kintamasis turi stiprų tiesinį sąryšį su `R.D.Spend` ir `Marketing.Spend` išlaidomis. Tiesinis sąryšis tarp `Profit` ir `Administration` išlaidų yra ganėtinai silpnas. `R.D.Spend` kintamasis silpnai koreliuoja su `Administration` kintamuoju, tačiau turi gan stiprią tiesinę priklausomybę su `Marketing.Spend`. `Administration` su `Marketing.Spend` beveik nekoreliuoja.

Duomenų tarpusavio ryšį galime pavaizduoti sklaidos diagrama. Grafiškai pavaizduojame `Profit` sąryšį su kiekvienu kintamuoju.



Duomenų sklaidos diagramos suteikia analogiškas įžvalgas kaip ir koreliacijos matrica – `Profit` turi stiprų tiesinį sąryšį su `R.D.Spend`, šiek tiek silpnesnį sąryšį su `Marketing.Spend` ir beveik visai nekoreliuoja su `Administration`.

1.5 Statistinės analizės išvados.

Atlikome išsamią analizę, kurios metu buvo ištirti duomenų empiriniai įverčiai, skirstiniai ir jų charakteristikos bei kintamųjų tarpusavio koreliacija. Remiantis gautais rezultatais, galima padaryti dvi išvadas:

i) duomenys yra paimti iš normaliosios populiacijos;
ii) pastebėti tam tikri tiesiniai ryšiai tarp kintamųjų.

Šios išvados yra naudingos kuriant tiesinį daugialypės regresijos modelį, kuris leis prognozuoti startuolio pelną remiantis įvairiomis išlaidomis.

2. Daugialypės regresijos modelis

Antrasis statistinio tyrimo tikslas yra sukurti daugialypės regresijos modelį, kuris galėtų tiksliai prognozuoti naujų startuolių pelną. Tam yra sukuriamas daugialypės regresijos modelis `linear_model`, kurį sieksime koreguoti taip, jog jis mums pateiktų geriausius rezultatus.

2.1 Tiesinio modelio prielaidos

Prieš įvertinant modelį, būtina įsitikinti, kad mūsų duomenys atitinka tiesinio modelio prielaidas. Todėl šioje dalyje patikrinsime pagrindines tiesinės regresijos modelio prielaidas. Tik įsitikinę, kad mūsų duomenys atitinka šias prielaidas, galėsime užtikrinti, jog sukurtas patikimas tiesinės regresijos modelis ir jo prognozės yra racionalios.

Tiesinio modelio prielaidos:

a) Liekamosios paklaidos turi būti normaliosios. Tai ekvivalentu, kad prognozuojamas kintamasis yra normalusis.
b) Regresoriai neturi stipriai koreliuoti. Kitaip gali iškilti multikolinearumo problema.
c) Duomenyse neturi būti išskirčių.
d) Duomenys turi būti homoskedastiški.

2.1.1 Pirmoji tiesinio modelio prielaida

Pirmoji tiesinio modelio prielaida teigia, jog liekamosios paklaidos turi turėti normalųjį skirstinį. Ši prielaida yra ekvivalenti prognozuojamo kintamojo normalumui. `Profit` yra mūsų priklausomas kintamasis, kurį norime prognozuoti. Šio kintamojo normalumui tirti naudojome histogramą, teorinį tankio grafiką, kvantilių grafiką bei Šapiro-Vilko testą. Padarėme išvadą, jog kintamasis yra paimtas iš normaliosios populiacijos. Tai reiškia, jog ir liekamosios paklaidos yra normaliosios. Taigi, pirmoji tiesinio modelio prielaida yra tenkinama.

Galime dar karta įsitikinti, jog liekamosios paklaidos turi normalųjį skirstinį. Tam naudojame Šapiro-Vilko kriterijų.

```
> shapiro.test(linear_model$residuals)

        Shapiro-Wilk normality test
```

```
data:  linear_model$residuals
W = 0.96114, p-value = 0.127
```

Gauta Šapiro-Vilko kriterijaus `p-value` lygi 0.127 > 0.05. Vadinasi, hipotezė, jog liekamosios paklaidos yra iš normaliosios populiacijos yra teisinga.

2.1.2 Antroji tiesinio modelio prielaida

Antroji prielaida sako, jog regresoriai tarpusavyje neturi stipriai koreliuoti. Tirdami duomenų tarpusavio priklausomybę nustatėme, jog `R.D.Spend` silpnai koreliuoja su `Administration` išlaidomis, tačiau su `Marketing.Spend` koreliuoja stipriai. Taip pat nustatėme, jog `Administration` su `Marketing.Spend` beveik nekoreliuoja. Tikslesniam duomenų multikolinearumo įvertinimui pasinaudosime dispersijos mazejimo daugikliu `vif`. `vif` koeficientas skaiciuojamas kiekvienam regresoriui. Jei `VIF` koeficientas > 4, tuomet multikolinearumas tarp tų kintamųjų yra.

Skaičiuojame dispersijos mažėjimo daugiklį `vif` kiekvienam regresoriui

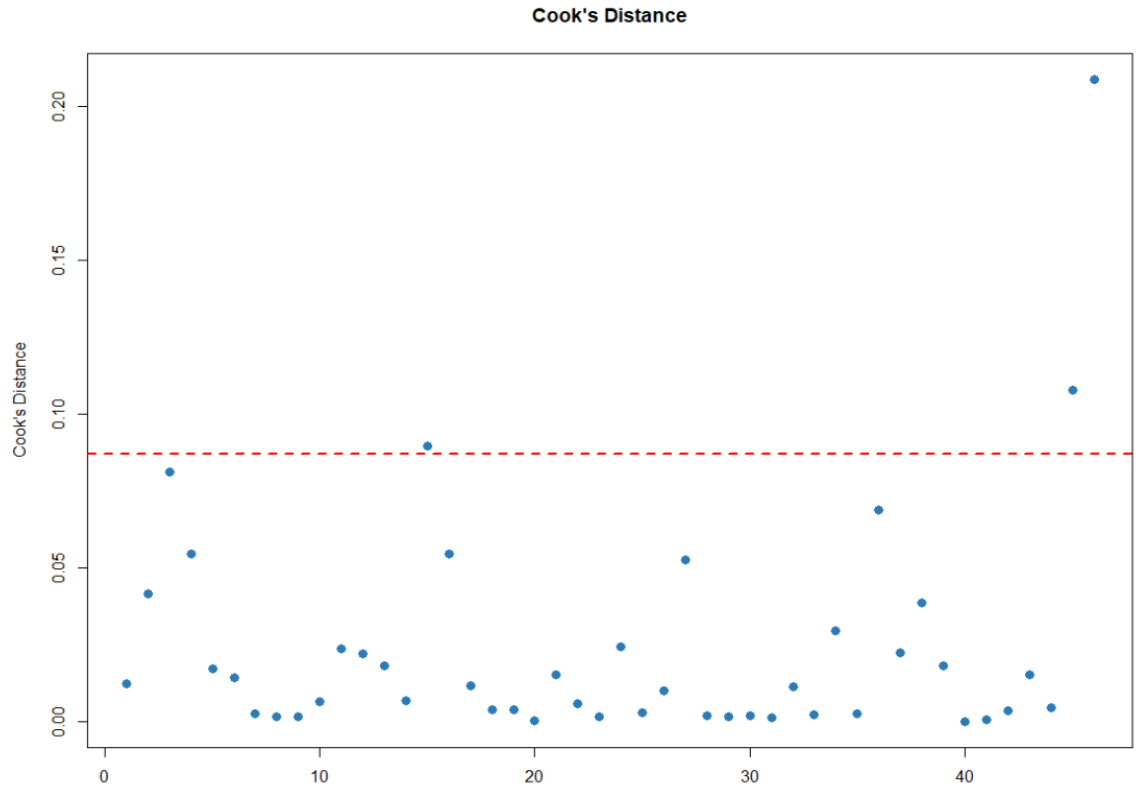
```
> vif(linear_model)
      data$R.D.Spend  data$Administration data$Marketing.Spend
            2.315286             1.162451             2.242803
```

Gauname, jog dispersijos mažėjimo daugikliai mažesni už 4 kiekvienam kintamajam, todėl galime teigti, jog multikolinearumo tarp regresorių nėra.

2.1.3 Trečioji tiesinio modelio prielaida

Trečioji prielaida teigia, jog duomenyse negali būti išskirčių. Tai nustatyti galime naudojant Kuko matą. Šis matas įvertina, kaip pasikeičia bendras modelio koeficiento pokytis pašalinus stebinį. Jis apskaičiuojamas kiekvienam stebiniui. Stebinys yra laikomas išskirtimi, jeigu to stebinio Kuko mato reikšmė yra didesnė negu 4/n, kur n - tiriamųjų duomenų imties dydis.

Nubrėžiame grafiką, kuris nurodo, kurių stebinių Kuko mato reikšmė viršija 4/n



Taigi, įvertinę Kuko mato reikšmes gavome, jog trys stebiniai yra išskirtys. Pašalinus šias išskirtis duomenys jau tenkins trečiąją tiesinio modelio prielaidą.

2.1.4 Ketvirtoji tiesinio modelio prielaida

Ketvirtoji prielaida tvirtina, jog duomenys turi būti homoskedastiški. Šią prielaidą galime patikrinti naudodami Breušo-Pagano kriterijų. Nulinė Breušo-Pagano kriterijaus hipotezė teigia, jog duomenys yra homoskedastiški.

```
> bptest(linear_model)

        studentized Breusch-Pagan test

data:  linear_model
BP = 1.077, df = 3, p-value = 0.7826
```

Gauta `p-value` 0.783 > 0.05, todel nuline hipoteze galime priimti. Vadinasi, duomenys yra homoskedastiški ir ketvirtoji tiesinio modelio prielaida yra tenkinama.

2.2 Tiesinio modelio tikslumo įvertinimas

Atnaujiname daugialypės regresijos modelį `linear_model_updated`, naudodami duomenis be išskirčių.

Modelio tikslumas matuojamas `R-squared` rodikliu. Šis dydis parodo kiek procentų prognozuojamo kintamojo elgesio lemia regresorių elgesys. Kuo `R-squared` reikšmė artimesnė 1, tuo geriau modelis aprašo duomenis.

`Adjusted R-squared` rodiklis atsižvelgia į duomenų imties dydį ir nepriklausomų kintamųjų modelyje skaičių, todėl geriau parodo modelio tinkamumą negu `R-squared` rodiklis.

T(Stjudento) kriterijai apskaičiuojami atskiriems regresoriams. Jie padeda nuspręsti, ar atitinkamas regresorius šalintinas iš modelio. Jeigu atitinkamo kriterijaus `p-value` mažesnė už 0.05, tariama, kad regresorius yra statistiškai reikšmingas ir (dažniausiai) modelyje yra paliekamas.

```
> summary(linear_model_updated)

Call:
lm(formula = data_cut$Profit ~ data_cut$R.D.Spend + data_cut$Administration +
    data_cut$Marketing.Spend)

Residuals:
     Min       1Q   Median       3Q      Max
-16094.9  -5070.9   -885.7   4892.5  12090.8

Coefficients:
                          Estimate  Std. Error t value             Pr(>|t|)
(Intercept)              54936.24788  5567.05225   9.868     0.00000000000372 ***
data_cut$R.D.Spend           0.76028     0.04041  18.814 < 0.0000000000000002 ***
data_cut$Administration     -0.03619     0.04124  -0.878               0.3855
data_cut$Marketing.Spend     0.03226     0.01513   2.132               0.0394 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6701 on 39 degrees of freedom
Multiple R-squared:  0.9653,   Adjusted R-squared:  0.9626
F-statistic: 361.3 on 3 and 39 DF,  p-value: < 0.00000000000000022
```

Atlikę regresijos modelio tikslumo įvertinimą gavome, jog `Adjusted R-squared` rodiklis lygus 0.9626. Vadinasi, tiesinės daugialypės regresijos modelis puikiai aprašo duomenis. Galime pastebėti, jog `Administration` regresoriaus T(Stjudento) kriterijaus `p-value` lygi 0.386 > 0.05. Vadinasi, daugiklis prie regresoriaus yra statistiškai nereikšmingas, todėl šį regresorių galime pašalinti.

Atnaujiname tiesinės daugialypės regresijos modelį `linear_model_final` pašalinę statistiškai nereikšmingą regresorių.

```
> summary(linear_model_final)

Call:
lm(formula = data_cut$Profit ~ data_cut$R.D.Spend + data_cut$Marketing.Spend)

Residuals:
     Min       1Q   Median       3Q      Max
-15833.0  -4116.4   -795.2   3835.6  11705.4

Coefficients:
                          Estimate  Std. Error t value            Pr(>|t|)
(Intercept)              50564.63428  2479.64916  20.392 <0.0000000000000002 *
**
data_cut$R.D.Spend           0.74684     0.03729  20.027 <0.0000000000000002 *
**
data_cut$Marketing.Spend     0.03688     0.01414   2.608               0.0128 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6681 on 40 degrees of freedom
Multiple R-squared:  0.9646,   Adjusted R-squared:  0.9628
F-statistic: 544.6 on 2 and 40 DF,  p-value: < 0.00000000000000022
```

Atlikę regresijos modelio tikslumo įvertinimą gavome, jog `Adjusted R-squared` rodiklis lygus 0.9628, t.y. rodiklis pakito labai nežymiai. Vadinasi, šis tiesinės daugialypės regresijos modelis taip pat puikiai aprašo duomenis. Visi sukurto modelio `linear_model_final` regresoriai yra statistiškai reikšmingi. Taip pat, duomenys, kurie buvo naudojami kuriant modelį, atitinka tiesinio modelio prielaidas.Remiantis turimais duomenimis ir atliktais statistiniais analizės metodais, galime užtikrinti, jog sukurtas patikimas tiesinės regresijos modelis. Modelis leidžia tiksliai prognozuoti startuolio pelną atsižvelgiant į įvairias įmonės išlaidas.

2.3 Tiesinio daugialypio modelio interpretacija ir pritaikymas

2.3.1 Daugialypio modelio interpretacija

Turėdami daugialypės regresijos modelį galime sužinoti regresijos funkcijos koeficientus.

```
> linear_model_final$coefficients
            (Intercept)       data_cut$R.D.Spend data_cut$Marketing.Spend
          50564.63427626              0.74684345               0.03687953
```

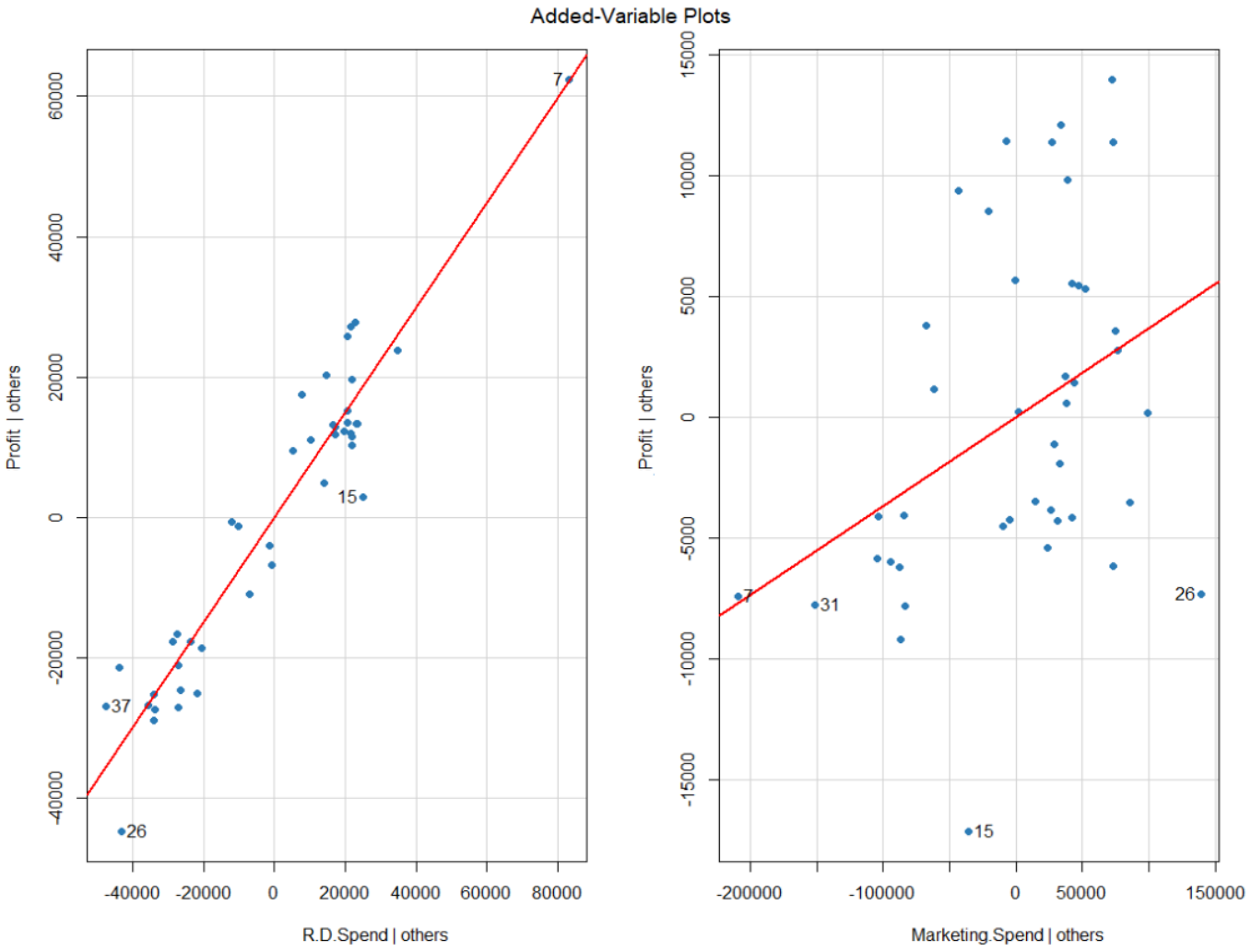
Matome, jog koeficientas prie `R.D.Spend` yra lygus 0.747, o koeficientas prie `Marketing.Spend` lygus 0.037. Galima pamanyti, jog kintamasis `R.D.Spend` yra žymiai labiau statistiškai reikšmingas ir modeliui užtektų tik šio regresoriaus. Tam, kad regresoriai rodytų tikslius duomenis, juos reikia standartizuoti.

```
> lm.beta(linear_model_final)
      data_cut$R.D.Spend data_cut$Marketing.Spend
               0.8925451                0.1162166
```

Gauti modelio su standartizuotais kintamaisiais koeficientai. Galima pastebėti, jog kintamasis `R.D.Spend` iš tikrųjų yra beveik 8 kartus statistiškai reikšmingesnis negu `Marketing.Spend`, t.y. santykis yra daug mažesnis negu nestandartizuotų kintamųjų.

Turint daugialypės regresijos modelį, jo rezultatus sunku pateikti vizualiai. Daugialypės regresijos grafiniam vaizdavimui yra naudojamas grafikas, sukuriamas su funkcija avPlots. Šis grafikas yra naudingas vizualinis įrankis, padedantis suprasti kiekvieno regresoriaus poveikį priklausomajam kintamajam.

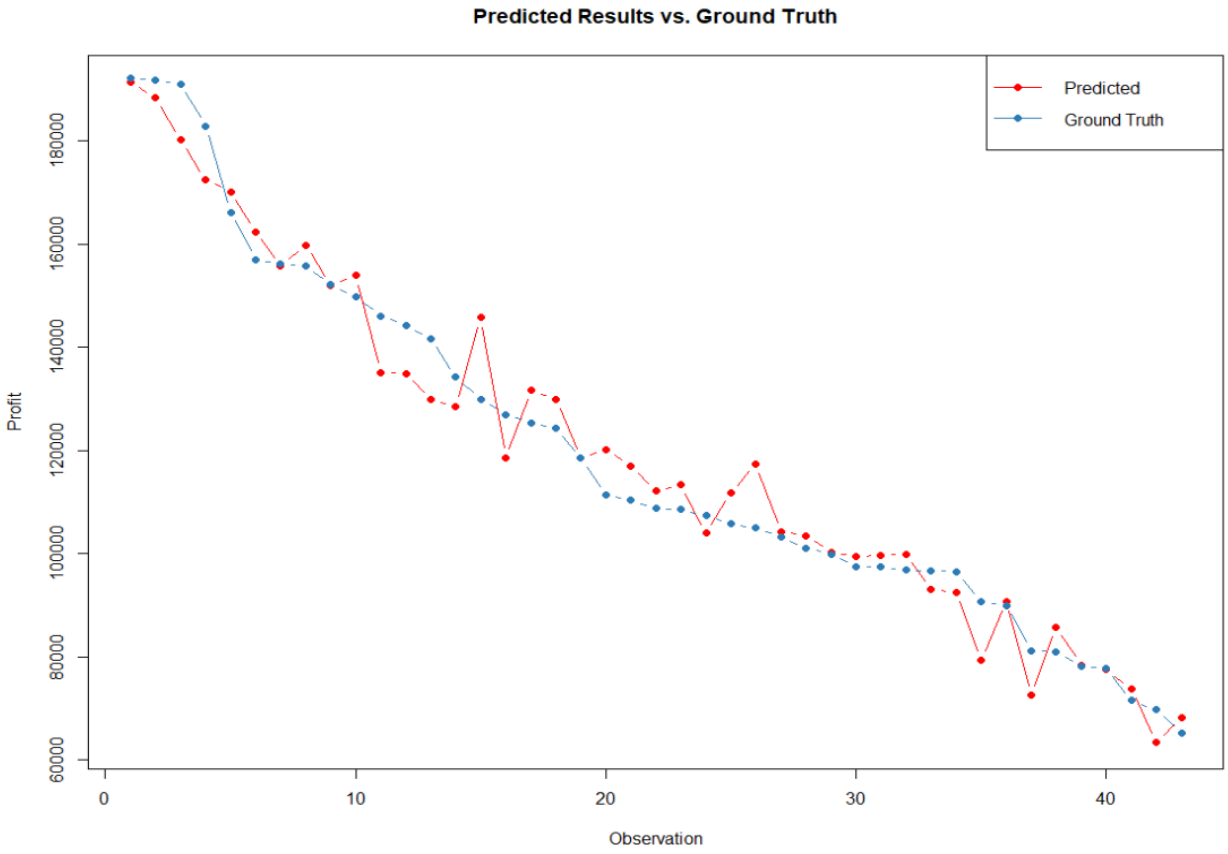
Kiekvienas regresorius atvaizduojamas atskirame grafike. Atskiri grafikai parodo regresoriaus ir priklausomo kintamojo santykį, kuomet kiti regresoriai yra laikomi konstantomis. Tai padeda vizualiai nustatyti kiekvieno regresoriaus ir priklausomo kintamojo priklausomybę.



2.3.2 Regresinio modelio pritaikymas

Turėdami tiesinės regresijos modelį, kuris leidžia tiksliai prognozuoti startuolio pelną atsižvelgiant į įvairias įmonės išlaidas, galime jį išbandyti.

Nubrėšime stebimų ir prognozuojamų reikšmių grafikus. Taip palyginsime modelio tiklsumą.



Matome, jog modelio `linear_model_final` prognozuojamos `Profit` reikšmės labai nedaug skiriasi nuo tikrųjų `Profit` reikšmių. Tai dar kartą parodo, jog modelis `linear_model_final` gali tiksliai prognozuoti naujų startuolių pelną.

Išbandykime sukurtą daugialypės regresijos modelį su naujais duomenimis. Tam sukurkime kelis rinkinius hipotetinių duomenų `new_data`, kuriuos panaudosime pelno prognozei.

```
> new_data
  R.D.Spend Administration Marketing.Spend
1     25000          120000          445350
2     73540          122500          238650
3    108000           67400          104560
```

```
> predict(linear_model_final, newdata = new_data)
       1        2        3
 85660.02 114288.80 135079.85
```

Hipotetinių duomenų prognozės mums parodo logiškai pagrįstus įverčius.

Išvados

Tyrimo tikslas buvo ištirti startuolių pelno ir MTEP (mokslinių tyrimų ir eksperimentinės plėtros) išlaidų, administravimo išlaidų ir rinkodaros išlaidų duomenis bei jų tarpusavio sąryšį ir sukurti tiesinės regresijos modelį, kuris tiksliai numato startuolio pelną pagal atitinkamas išlaidas. Atlikus statistinį tyrimą, galima padaryti šias išvadas:

- Išnagrinėjus duomenis, galima daryti išvadą, kad duomenys atitinka normaliosios populiacijos sąlygas. Tai leidžia mums taikyti statistinę analizę ir modeliavimo metodus.
- Pastebėti tam tikri tiesiniai ryšiai tarp kintamųjų. Tai reiškia, kad pelnas gali būti tiksliai prognozuojamas, remiantis įmonės veiklos išlaidomis, tokiomis kaip MTEP išlaidos, administravimo išlaidos ir rinkodaros išlaidos.
- Sukurtas daugialypės regresijos modelis, kuris gerai aprašo duomenis ir leidžia tiksliai prognozuoti startuolių pelną, atsižvelgiant į įmonės veiklos išlaidas.

Šios išvados rodo, kad atlikus statistinį tyrimą ir sukūrus tinkamą modelį, galima gauti vertingų žinių investuotojams ir suinteresuotiems asmenims, siekiantiems įvertinti startuolių finansines perspektyvas.