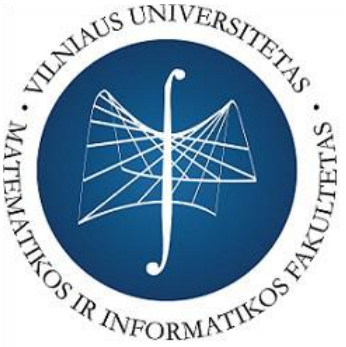


VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

DUOMENŲ MOKSLAS. BAKALAURAS

3 Laboratorinis darbas

Kvantilių regresija

Ataskaita

Užduotį atliko: Ugnius Vilimas

Rytis Baltaduonis

Justinas Pipiras

Vilnius 2024

Turinys

Įvadas

Kvantilių regresija yra pagrįsta kvantilių tiesinės regresijos modeliais. Šie modeliai skiriasi nuo įprastinių tiesinės regresijos modelių tuo, kad jie numato ne tiesiog vidurkį, o tam tikrą kvantilį (pavyzdžiui, 25% arba 75%) kaip atsako kintamojo. Šis metodas yra naudingas, kai turime duomenis, kuriuose yra ryšių, kurių statistinė struktūra kinta skirtinguose kvantiliuose.

Darbo tikslas

- Atlikti kvantilių regresijos modelio analizę, mūsų pasirinktam duomenų rinkiniui.

Uždaviniai

- Pasirinkti duomenų rinkinį iš prieinamo duomenų šaltinio.
- Susitvarkyti gautus duomenis.
- Atlikti pradinę duomenų analizę.
- Sukurti kvantilių regresijos modelį.
- Pateikti tyrimo rezultatus ir išvadas.

Laboratorinis darbas buvo atliekas su „R“ programine įranga.

Duomenys

Duomenys yra apie krepšininkų sezono statistiką nuo 1999 iki 2020. Duomenys imti iš „Kaggle“. Mūsų duomenų rinkinį sudaro 53949 eilučių ir 34 stulpelių. Mes pasirinkome tokius kintamuosius:

- GP – sužaistų rungtynių kiekis per sezoną.
- MIN – sužaistų minučių kiekis per sezoną.
- FGA – bandymų mesti į krepšį per sezoną.
- X3PA – bandymų mesti į krepšį iš trijų taškų zonos per sezoną.
- FTA – bandymų mesti į krepšį nuo baudų metimo linijos per sezoną.
- TOV – padarytų klaidų skaičius per sezoną.
- PF – padarytų pražangų skaičius per sezoną.
- REB – atkovotų kamuolių skaičius per sezoną.
- AST – rezultatyvių perdavimų skaičius per sezoną.
- STL – perimtų kamuolių skaičius per sezoną.
- BLK – blokuotų metimų skaičius per sezoną.
- PTS – įmestų taškų skaičius per sezoną.

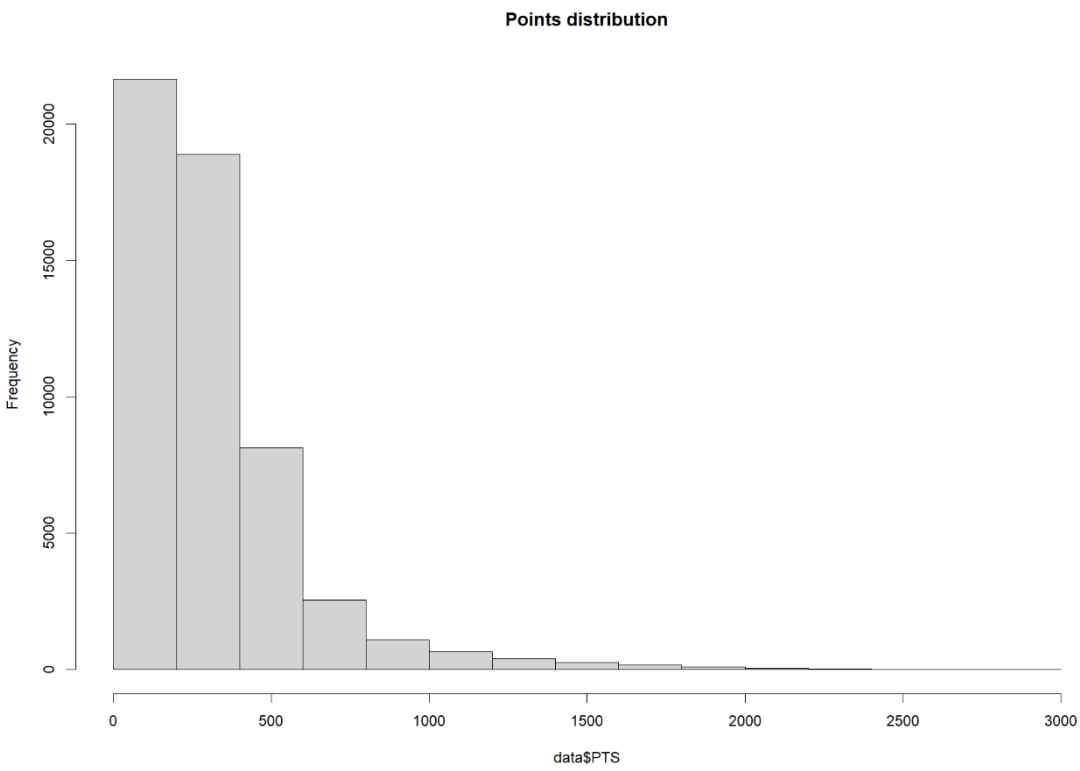
PTS buvo mūsų priklausomas kintamasis, kiti – nepriklausomi kintamieji. Visi kintamieji buvo kiekybiniai. Praleistų reikšmių pasirinktuose kintamuosiuose nebuvo, todėl naudojome visus duomenis. Mokymo ir testavimo duomenys buvo padalinti santykiu 80/20.

Analizė su „R“

Pradinė analizė

Sudarinėjant kvantilių regresijos modelį, didelių reikalavimų duomenims nėra, tam visiškai nėra svarbus priklausomo kintamojo skirstinys ir dispersija. Taip pat modelis nėra jautrus išskirtims, todėl pasirinkome jų net netikrinti. Žinome, kad kvantilių regresijos modelis veikia geriausiai, kai turime kuo daugiau duomenų.

Čia matome mūsų priklausomo kintamojo PTS pasiskirstymą grafiškai.



1 pav. įmestų taškų pasiskirstymas

Daugiausiai reikšmių matome tarp 0 ir 200. Taip pat matome, kad surinktų taškų skaičiui didėjant, mažėja dažnumas.

(2 Pav.) pavaizdavome priklausomojo kintamojo pagrindines charakteristikas, tokias kaip: minimali reikšmė (nulinis kvantilis), pirmas ir trečias kvantiliai, mediana (antras kvantilis), maksimali reikšmė (ketvirtas kvantilis) bei vidurkis.

```
> summary(data$PTS)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   134.0   247.0   311.2   399.0  2832.0
```

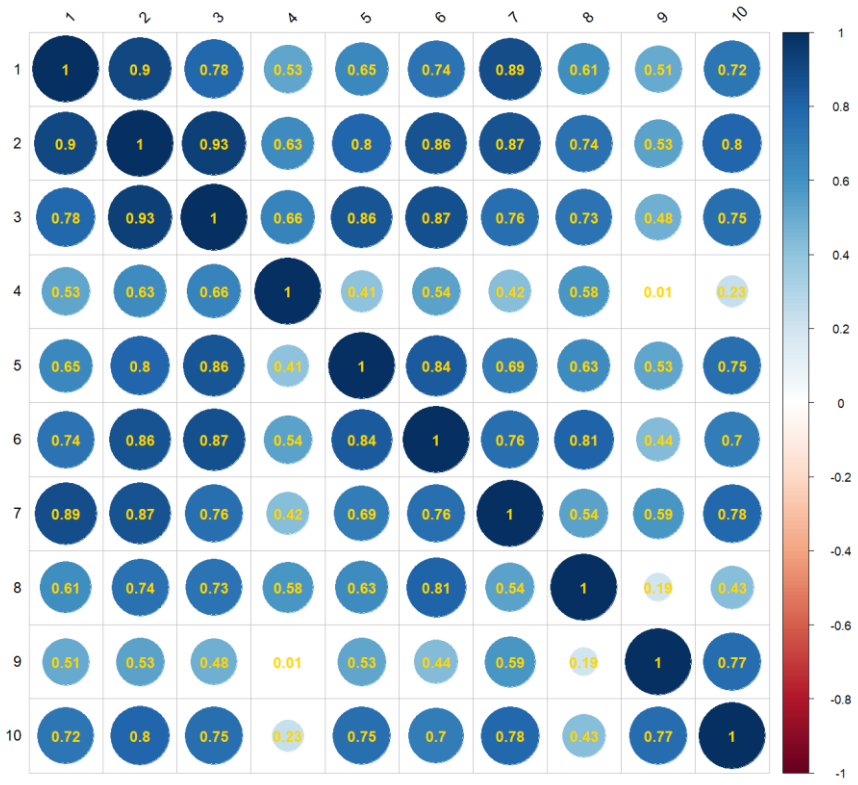
2 pav. priklausomojo kintamojo charakteristikos

(3 Pav.) pavaizdavome nepriklausomų kintamųjų pagrindines charakteristikas, tokias kaip: minimali reikšmė (nulinis kvantilis), pirmas ir trečias kvantiliai, mediana (antras kvantilis), maksimali reikšmė (ketvirtas kvantilis) bei vidurkis.

```
> summary(cbind(data$GP, data$MIN, data$FGA,data$X3PA, data$FTA, data$TOV, data$PF, data$AST, data$BLK, data$REB))
       V1              V2              V3              V4              V5              V6              V7       
 Min.   : 1.00   Min.   :   0.0   Min.   :   0.0   Min.   :   0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.0  
 1st Qu.:17.00   1st Qu.: 380.9   1st Qu.: 109.0   1st Qu.:  20.00   1st Qu.: 28.00   1st Qu.: 21.00   1st Qu.: 37.0  
 Median :29.00   Median : 663.0   Median : 196.0   Median :  61.00   Median : 56.00   Median : 39.00   Median : 64.0  
 Mean   :30.31   Mean   : 752.4   Mean   : 245.1   Mean   :  80.74   Mean   : 76.01   Mean   : 47.26   Mean   : 70.1  
 3rd Qu.:37.00   3rd Qu.: 954.0   3rd Qu.: 310.0   3rd Qu.: 117.00   3rd Qu.: 99.00   3rd Qu.: 63.00   3rd Qu.: 91.0  
 Max.   :85.00   Max.   :3485.0   Max.   :2173.0   Max.   :1028.00   Max.   :972.00   Max.   :464.00   Max.   :371.0  
       V8              V9              V10       
 Min.   :  0.00   Min.   :  0.00   Min.   :   0.0  
 1st Qu.: 20.00   1st Qu.:  1.00   1st Qu.:  50.0  
 Median : 41.00   Median :  4.00   Median :  93.0  
 Mean   : 62.79   Mean   : 10.49   Mean   : 124.9  
 3rd Qu.: 78.00   3rd Qu.: 12.00   3rd Qu.: 159.0  
 Max.   :925.00   Max.   :307.00   Max.   :1247.0  
```

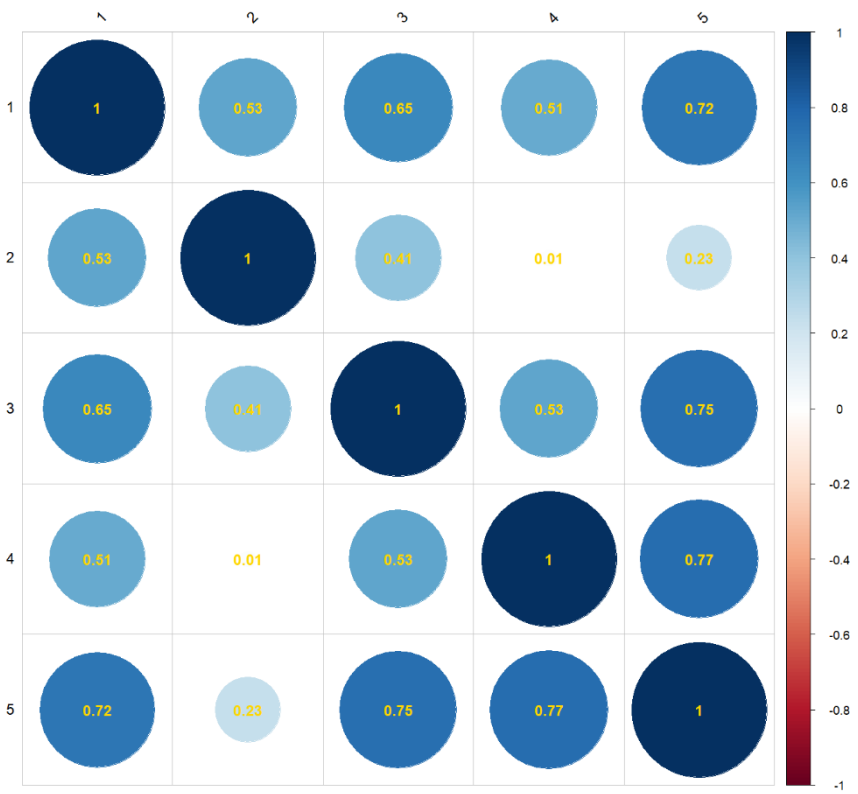
2 pav. nepriklausomų kintamųjų charakteristikos

(4 pav.) matome koreliacijos matricą. Dauguma kintamųjų gana stipriai koreliuoja, todėl nusistamėme, kad kintamųjų, kurių koreliacija didesnė už 0.8, į modelį nedėsime.



3 pav. kintamųjų koreliacijos matrica

(5 pav.) sudarome tik mums reikšmingų kintamųjų koreliacijos matricą. Į ją patenka tokie kiekybiniai kintamieji: GP, X3PA, FTA, BLK, REB.



4 pav. reikšmingų kintamųjų koreliacijos matrica

Modelis

Toliau tikrinsime su kuria τ reikšme modelis veikia geriausiai. Tikrinsime 3 τ reikšmes(0.25, 0.5, 0.75).

```
Call: rq(formula = PTS ~ GP + X3PA + FTA + BLK + REB, tau = 0.5, data = train_data)

tau: [1] 0.5

Coefficients:
            Value     Std. Error t value   Pr(>|t|)
(Intercept)  -7.67265   0.29995  -25.57942   0.00000
GP            0.28751   0.03422    8.40249   0.00000
X3PA          1.06472   0.00721  147.76579   0.00000
FTA           1.95408   0.01270  153.87894   0.00000
BLK           0.24273   0.04865    4.98936   0.00000
REB           0.54616   0.01016   53.73244   0.00000
```

5 pav. modelis su τ = 0.5

(6 pav.) pavaizduotam modelyje visi likę kintamieji yra statistiškai reikšmingi, dėl to modelį su τ = 0.5 galėtume naudoti galutiniams spėjimams.

```
Call: rq(formula = PTS ~ GP + X3PA + FTA + BLK + REB, tau = 0.25, data = train_data)

tau: [1] 0.25

Coefficients:
            Value     Std. Error t value   Pr(>|t|)
(Intercept) -10.43388   0.35045  -29.77250   0.00000
GP            0.02948   0.03113    0.94723   0.34353
X3PA          1.04701   0.00658  159.18736   0.00000
FTA           1.79192   0.01214  147.65768   0.00000
BLK           0.21433   0.05107    4.19678   0.00003
REB           0.47103   0.00961   49.00192   0.00000
```

7 pav. modelis su τ = 0.25

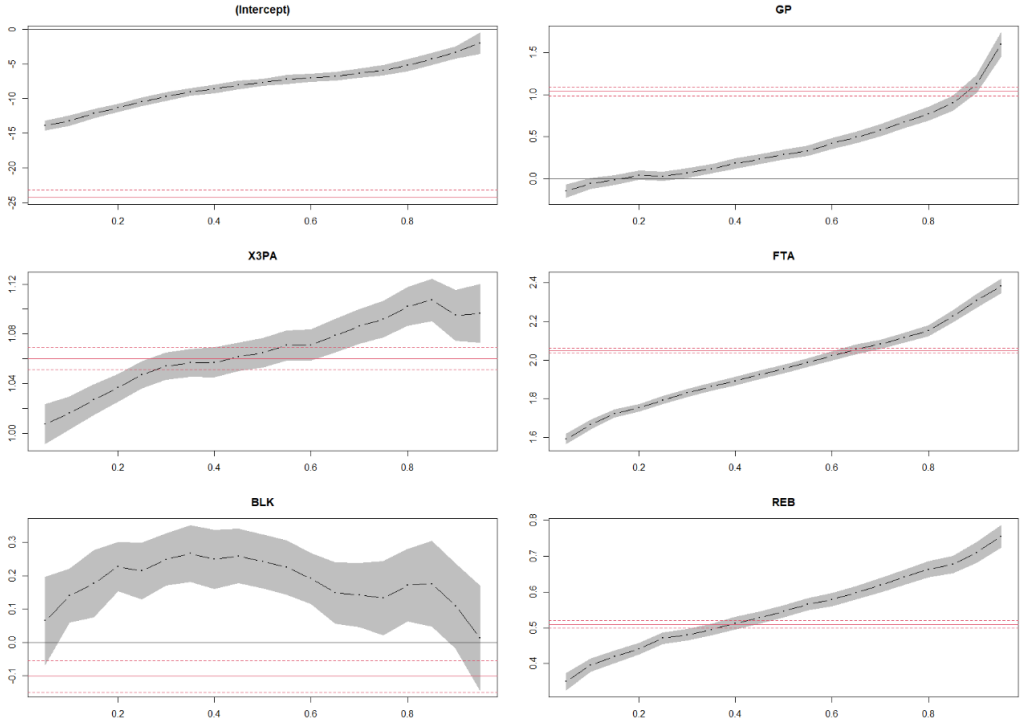
(7 pav.) gauname, kad nepriklausojomo gintamojo GP(sužaisti žaidimai) p reikšmė(0.343) yra didesnė už alfa(0.05) dėl to šio modelio galutiniams spėjimams nenaudosime.

```
Call: rq(formula = PTS ~ GP + X3PA + FTA + BLK + REB, tau = 0.75, data = train_data)

tau: [1] 0.75

Coefficients:
            Value     Std. Error t value   Pr(>|t|)
(Intercept)  -5.91835   0.41345  -14.31440   0.00000
GP            0.67744   0.04307   15.72993   0.00000
X3PA          1.09184   0.00879  124.21497   0.00000
FTA           2.11667   0.01499  141.23133   0.00000
BLK           0.13248   0.06686    1.98160   0.04753
REB           0.64155   0.01237   51.86336   0.00000
```

8 pav. modelis su τ = 0.75

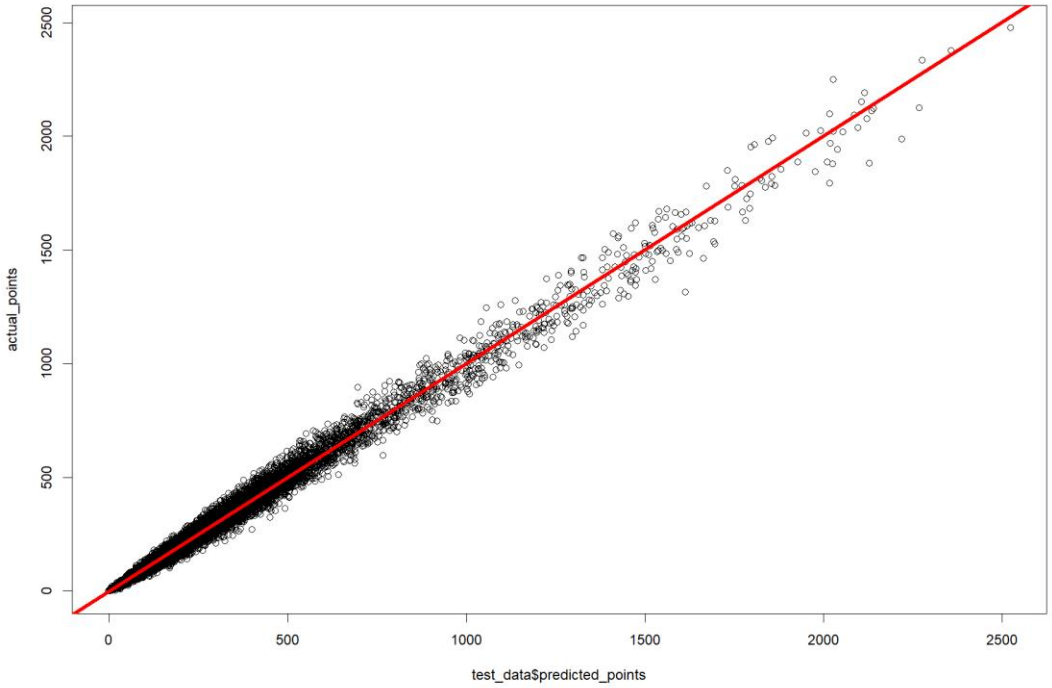
(8 pav.) gauname, kad nepriklausojomo gintamojo BLK(užblokuoti metimai) p reikšmė(0.0475) yra arti alfa reikšmės(0.05) dėl to šis modelis nėra toks geras kaip (6 pav.) su τ = 0.5 .

Kiekvienam kvantiliui apmokėme modelį (nuo 0,05 iki 0,95 su 0,05 intervalais). Iš pavaizduotų grafikų matome kaip kiekvienam kvantiliui keičiasi koeficientai (9 pav.)



6 pav. Koeficientų pokytis skirtinguose kvantiliuose.

Rezultatai



7 pav. tikrųjų ir spėjamų reikšmių palyginimas

(10 pav.) Modelis įmestų taškų kiekį prognozuoja ganėtinai gerai, truputi didesnis netikslumas atsiranda didėjant per sezoną įmestų tašku kiekiui, nes tokių duomenų yra mažiau.

Išvados

Susitvarkius duomenis, atlikus pradinę analizę, apmokėme pasirinktus kvantilių modelius su skirtingomis τ reikšmėmis. Rinkomės tikrinti τ reikšmes: 0.25, 0.5 ir 0.75. Būtent mūsų duomenims su krepšininkų statistika, geriausiai veikė τ = 0.5. Galutinis modelis veikė puikiai, nematėme jokių nutolusių reikšmių. Turime nuojautą, kad yra dar geresnių parametrų apie žaidėjus, kurie leistų dar geriau nustatyti įmestų taškų kiekį.

Šaltiniai

- https://www.kaggle.com/datasets/jacobbaruch/basketball-players-stats-per-season-49-leagues