MATEMATIKOS IR INFORMATIKOS
FAKULTETAS

KVANTILIŲ REGRESIJA

Laboratorinis darbas

Atliko: Simona Gelžinytė,
Laineda Morkytė,
Austėja Valeikaitė,
duomenų mokslas 3 k. 2gr.

Vilnius, 2023

TURINYS

1. ĮVADAS

1.1 Tyrimo tikslas

Pritaikyti kvantilių regresijos modelį pasirinktiems duomenims.

1.2 Tyrimo uždaviniai

- Atlikti pirminę duomenų analizę;
- Patikrinti modelio prielaidas;
- Sukonstruoti modelį;
- Įvertinti modelio tinkamumą;
- Pateikti gauto modelio interpretacijas;
- Apibendrinti gautus rezultatus, pateikti išvadas.

1.3 Duomenys ir programinė įranga

Duomenų rinkinys pasirinktas apie nėštumus 1960-1967 m. tarp San Francisko Rytų įlankos regiono moterų. Pateikti įvairūs požymiai apie nėščiąją, kurie prognozuoja naujagimio svorį. Priklausomas kintamasis – naujagimio svoris (uncijomis) ir 6 kovariantės:

- Nėštumo laikotarpis (dienomis);
- Ar pirmas nėštumas (kategorinis): 0 – pirmas, 1 – kitu atveju;
- Moters amžius;
- Moters svoris (uncijomis);
- Moters ūgis (coliais);
- Ar moteris rūko (kategorinis): 0 – nerūko, 1 – rūko;

Iš viso yra 1236 stebėjimai, pašalinus praleistas reikšmes stebėjimų liko 1174, duomenų nepatikimumo problemos neturime (praleistų reikšmių 5 %). Taip pat turime pakankamai duomenų, kad būtų galima taikyti kvantilių regresiją. Tyrimo metu naudota „R“ programinė įranga.

2. KVANTILIŲ REGRESIJOS MODELIS

Parinkus kvantilių regresijos modelį pereiname visus modelio parinkimo etapus:

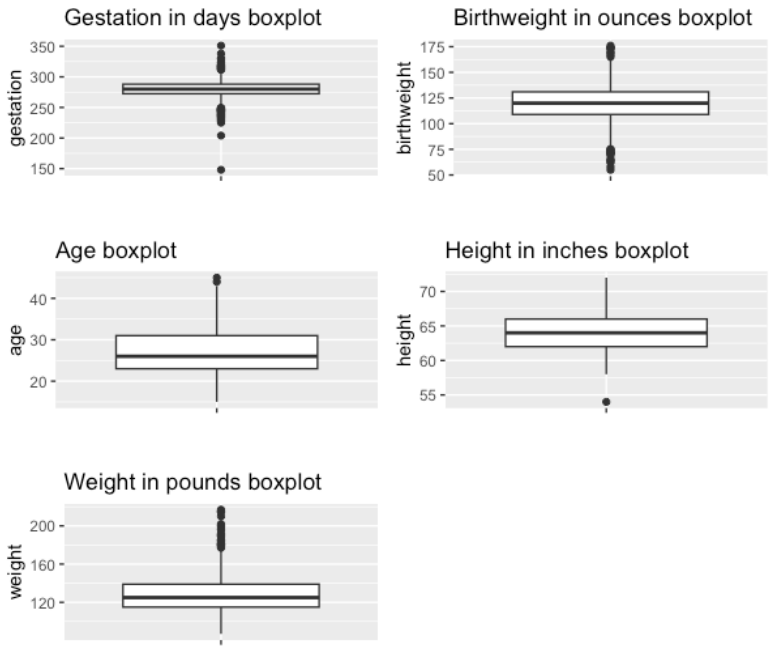
1. Pradinė analizė – vizualiai patikriname, ar nėra išskirčių;
2. Prielaidų tikrinimas – ar yra tiesinis sąryšis tarp priklausomo kintamojo ir kovariančių, multikolinearumo problema;
3. Reikšmingų kovariančių atranka;
4. Parametrų įvertinimas, interpretacija.

Duomenys buvo padalinti į mokymo ir testavimo aibes 70:30 santykiu. Pirminis kvantilių regresijos modelis atrodo taip:

$$Naujagimio\ svoris = \beta_0(\tau) + \beta_1(\tau) \cdot nėštumo\ laikotarpis + \beta_2(\tau) \cdot pirmas\ nėštumas + \beta_3(\tau) \cdot amžius + \beta_4(\tau) \cdot svoris + \beta_5(\tau) \cdot ūgis + \beta_6(\tau) \cdot rūko.$$

2.1 Kvantilių regresijos modelis naudojant R

Prieš sudarant kvantilių regresijos modelį, vizualiai pasižiūrime į duomenis. Iš stačiakampių diagramų matome, jog daugiausiai išskirčių turi nėštumo laikotarpis, kūdikio svoris ir nėščiosios svoris, tačiau labai išsiskiriančių reikšmių nėra (1 pav.). Iš duomenų aprašomosios statistikos (1 lentelė) matome, jog naujagimių svorio medianinė reikšmė – 120, mažiausias naujagimio svoris – 55 uncijos, o didžiausias – 176, nėštumo trukmės medianinė reikšmė – 280, nėščiosios amžius – 26, ūgis – 64, svoris – 125. Matome, jog net 598 (beveik 73%) moterims iš 823, tai buvo pirmasis nėštumas, o 327 (beveik 40%) moterys atsakė, jog yra rūkančios (2, 3 lentelės).



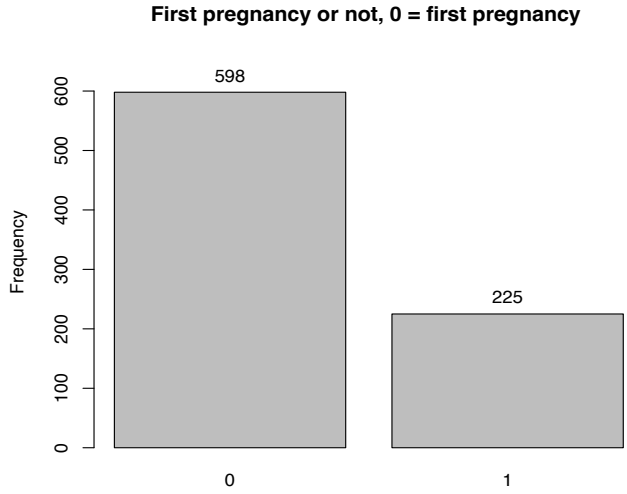
1 pav. Stačiakampių diagramų panelė

1 lentelė. Pradinė duomenų aprašomoji statistika

	vidurkis	**mediana**	**min**	**max**	Q_1	Q_3
Naujagimio svoris	119,6	120	55	176	109	131
Nėštumo laikotarpis	279	280	148	351	273	288
Amžius	27	26	15	45	23	31
Nėščiosios svoris	63,98	64	54	72	62	66
Nėščiosios ūgis	128,6	125	87	217	115	139

2 lentelė. Dažnių lentelė apie nėštumą

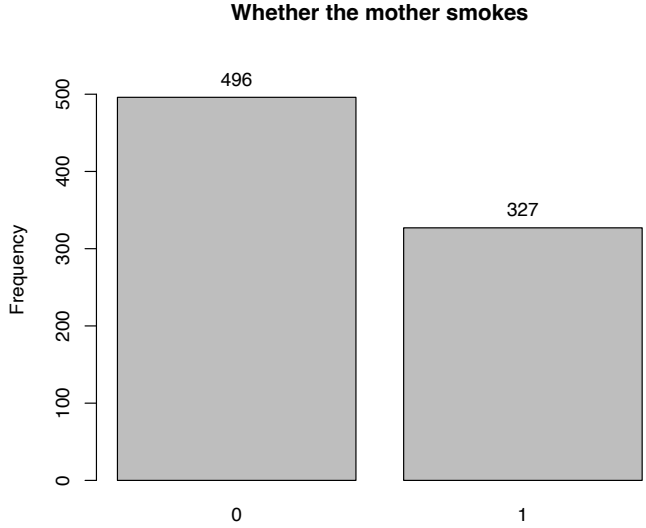
Pirmas nėštumas	**Ne pirmas nėštumas**
598	225



2 pav. Histograma apie nėštumą

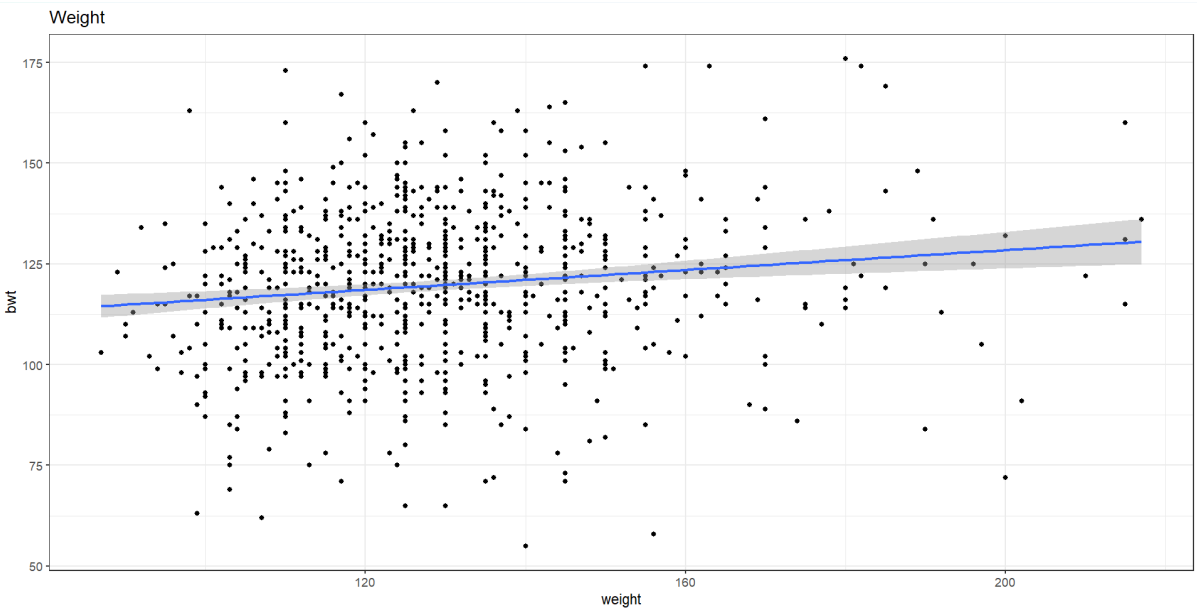
3 lentelė. Dažnių lentelė apie rūkymą

Rūko	Nerūko
327	496

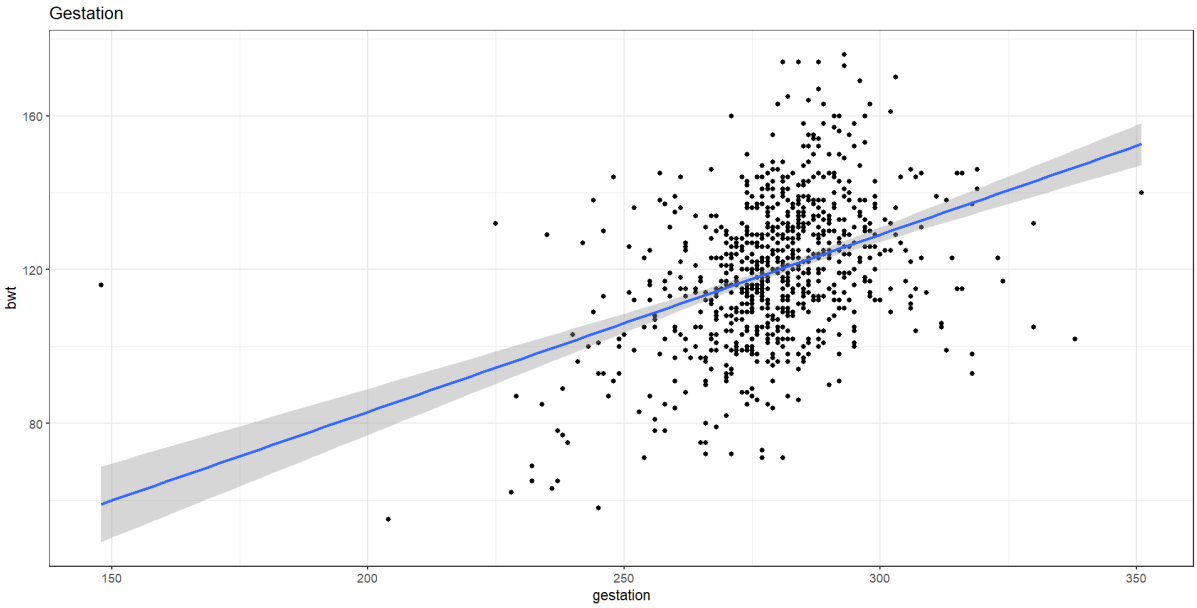


3 pav. Histograma apie rūkymą

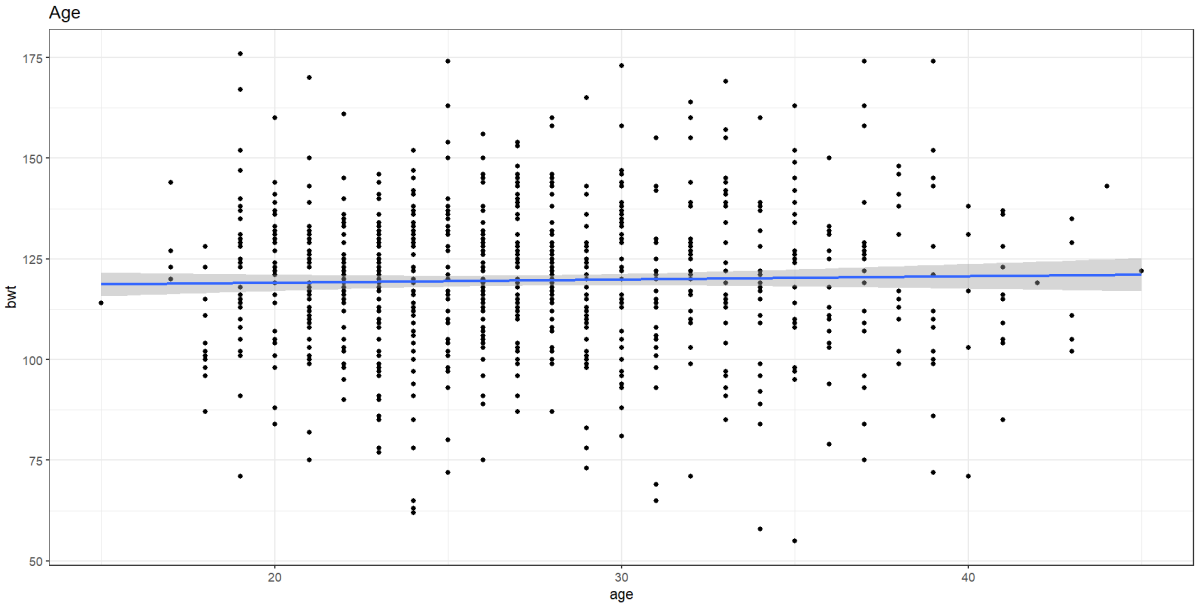
Atlikę pirminę duomenų analizę pritaikome kvantilių regresijos modelį ir tikriname prielaidas. Pirmiausia, tariame, jog atsitiktinių paklaidų kvantiliai lygus nuliui. Toliau turime patikrinti tiesiškumą – ar yra tiesinis sąryšis tarp priklausomo kintamojo ir kiekvienos kovariantės. Tam brėžiame grafikus.



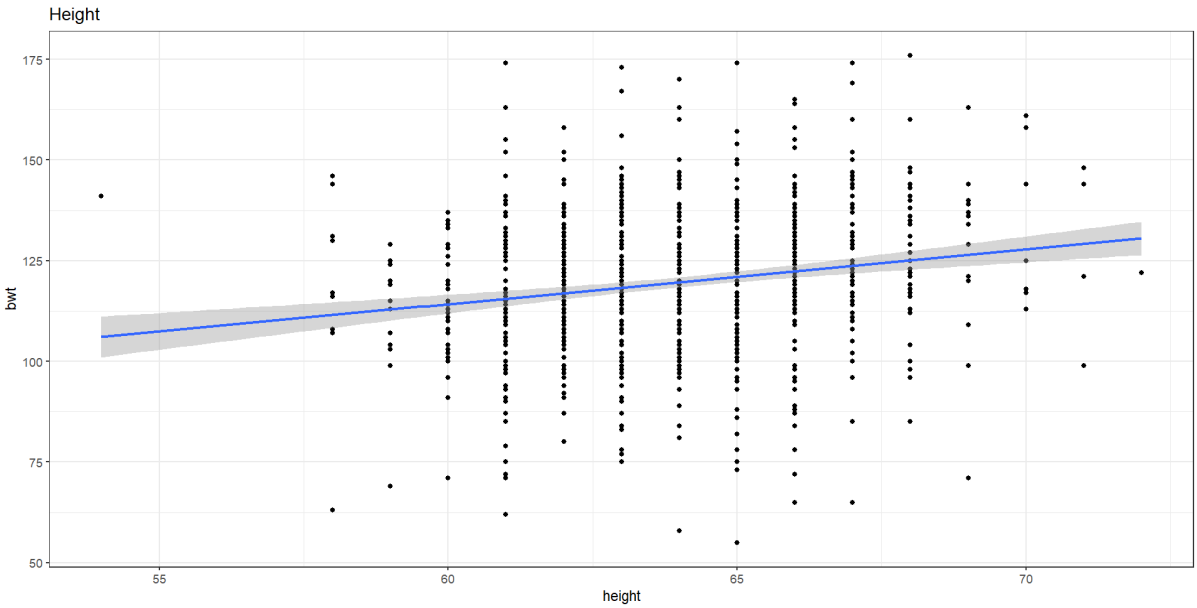
4 pav. Tiesinis sąryšis tarp kūdikių svorių ir motinų svorių



5 pav. Tiesinis sąryšis tarp kūdikių svorių ir nėštumo laikotarpio



6 pav. Tiesinis sąryšis tarp kūdikių svorių ir motinų amžiaus



7 pav. Tiesinis sąryšis tarp kūdikių svorių ir motinų ūgio

Galime pastebėti, jog tiesiškumo sąlyga išpildyta beveik visoms kovariantėms, bet tiesinį sąryšį tarp naujagimių svorių ir nėščiųjų amžiaus sunku įžvelgti. Nėščiųjų amžiaus kovariantei atlikome logaritminį transformavimą, tačiau jis nepadėjo. Taip pat matome, jog tiesinis modelis nebūtų tinkamas, nes jis blogai aprašo duomenis, todėl reiktų rinktis kvantilių regresiją.

Taip pat turime patikrinti ar neturime multikolinearumo problemos. Tikrinsime modelius, su visomis kovariantėmis bei atskirai su kiekviena kovariante ir lyginsime jų

koeficientų ženklus. Koeficientų ženklų netikrinsime pirmojo nėštumo ir rūkymo kovariantėms, nes tai kategoriniai kintamieji.

4 lentelė. Modelio, su visomis kovariantėmis, koeficientai

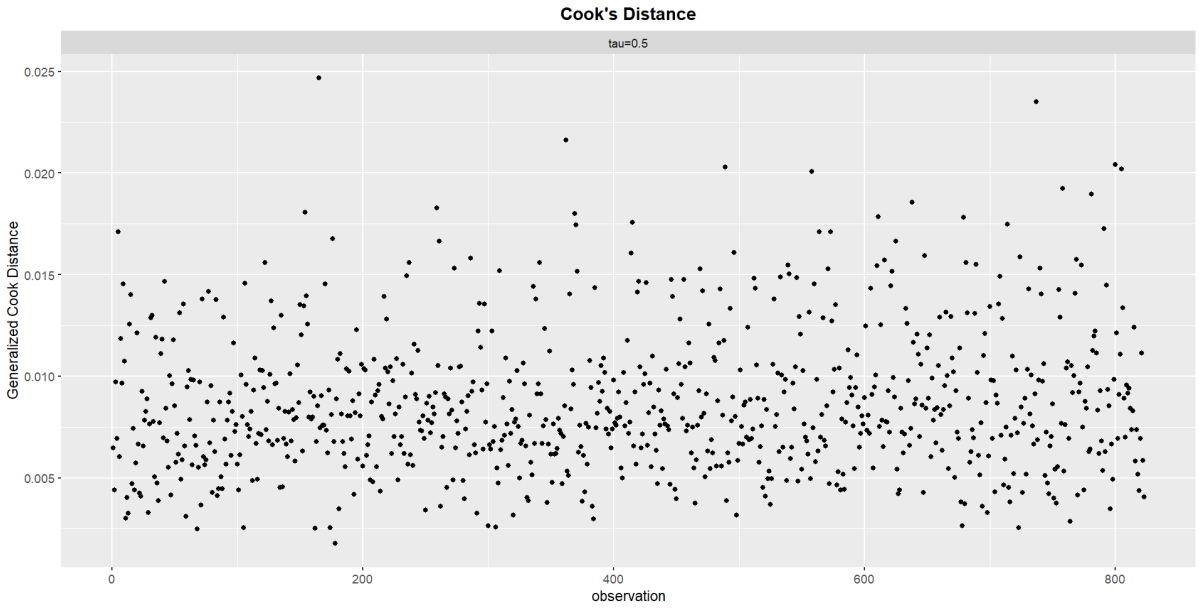
Kovariantės	Nėštumo laikotarpis	Moters amžius	Moters svoris	Moters ūgis
Koeficientai	0.4868	- 0.0015	0.0684	1.1192

5 lentelė. Modelio, su atskiromis kovariantėmis, koeficientai

Kovariantės	Nėštumo laikotarpis	Moters amžius	Moters svoris	Moters ūgis
Koeficientai	0.5348	1.5067e-15	0.1236	1.5

Kaip matome visi ženklai sutampa, išskyrus amžiaus. Tai reiškia, jog multikolinearumo problemą gali sukelti būtent ši kovariantė. Ją fiksuojame ir tikrinsime, ar atlikus pažingsninę regresiją amžiaus kovariantė išliks reikšminga, jei ne – multikolinearumo problema bus išspręsta.

Nors kvantilių regresija nėra jautri išskirtims, labai išsiskiriančios reikšmės vis tiek gali turėti įtakos gautiems rezultatams, todėl naudojant Kuko matą tai patikrinsime.



8 pav. Kuko išskirčių grafikas

Iš pateikto grafiko (8 pav.) matome, jog nei vienas stebėjimas neviršija 1, todėl duomenyse išskirčių nėra.

Kadangi modelio prielaidos patenkintos, tikriname, kurios kovariantės modelyje yra statistiškai reikšmingos (0.5 kvantilis). Tikriname hipotezę:

$$\begin{cases} H_0: \beta_j = 0 \\ H_A: \beta_j \neq 0 \end{cases}, kur\ j = 1, 2, \dots, 6.$$

```
Call: rq(formula = bwt ~ ., tau = 0.5, data = df)

tau: [1] 0.5

Coefficients:

            coefficients lower bd   upper bd

(Intercept)  -91.81141   -121.90946  -47.97889

gestation      0.47582      0.38651    0.57437

parity1       -3.72289     -6.49735   -1.27540

age           -0.05608     -0.28102    0.13266

height         1.22653      0.58903    1.68455

weight         0.04618     -0.00583    0.11250

smoke1        -8.02127    -10.54651   -5.57363
```

Gauta modelio santrauka pateikia parametrų įverčius ir pasikliovimo intervalus, tačiau nežinome kurios kovariantės yra reikšmingos. Taigi, toliau pritaikome pažingsninę regresiją (*both*, *backward* ir *forward*).

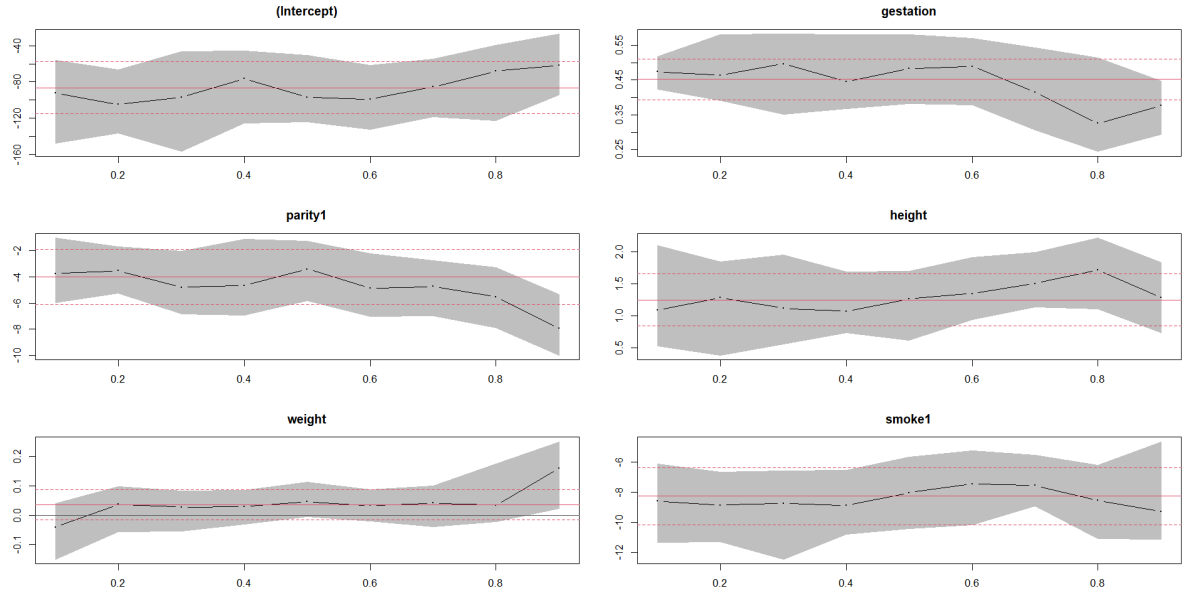
```
Start:  AIC=6955.95

bwt ~ gestation + parity + age + height + weight + smoke

            Df    AIC

- age        1 6954.3

<none>         6956.0

- weight     1 6956.5

- parity     1 6963.7

- height     1 6977.2

- smoke      1 7003.7
```

```
- gestation  1 7064.1

Step:  AIC=6954.33

bwt ~ gestation + parity + height + weight + smoke

            Df    AIC

<none>         6954.3

- weight     1 6954.6

+ age        1 6956.0

- parity     1 6961.7

- height     1 6976.8

- smoke      1 7003.2

- gestation  1 7062.6

Call:

rq(formula = bwt ~ gestation + parity + height + weight + smoke,

    tau = 0.5, data = df)

Coefficients:

 (Intercept)    gestation      parity1       height       weight
smoke1

-97.05911950   0.48238994  -3.42515723   1.25849057   0.04528302
-8.01761006

Degrees of freedom: 823 total; 817 residual
```

Pritaikius *both* ir *backward* pažingsninę regresiją, gavome, jog visos kovariantės yra reikšmingos išskyrus amžių. Abiem atvejais AIC koeficientai sutapo – 6954.33. *Forward* pažingsinė regresija paliko visas kovariantes, bet jos AIC buvo didesnis – 6955.95. Taip pat žinome, jog amžiaus kovariantė gali sukelti multikolinearumo problemą, todėl naudosime *both* pažingsninės regresijos atrinktas kovariantes.

Žemiau pateikta grafikų panelė (9 pav.) parodo kaip keičiasi kiekvienos kovariantės įvertintų koeficientų reikšmės skirtingiems kvantiliams. Matome, kad jei šiems duomenims būtų naudojamas vidurkis, daugeliui kovariančių modelis būtų aprašytas pakankamai gerai. Vis

dėlto, nėštumo laikotarpio, pirmojo nėštumo ir svorio kovariančių koeficientų reikšmės prie skirtingų kvantilių skiriasi labiau ir naudojant vidurkį modelis būtų aprašytas prasčiau. Matome, jog ilgėjant nėštumo laikotarpiui naujagimio svoris didėja (vidutiniškai 0.45 uncijomis), tačiau prie 0.8 ir 0.9 kvantilių, svoris padidėja ne tiek daug kiek prie 0.1 ar 0.5 kvantilių. Jei nėštumas moteriai nebuvo pirmas, tai kūdikio svoris bus mažesnis (vidutiniškai 4 uncijomis mažesnis). Kuo nėščioji yra aukštesnė ir daugiau sveria, tuo naujagimio svoris bus didesnis (vidutiniškai 1.25 ir 0.05 uncijos). Naujagimis taip pat svers mažiau (vidutiniškai 8 uncijomis mažiau), jei moteris yra rūkanti.



9 pav. Koeficientų reikšmės skirtingiems kvantiliams

Toliau, detaliau panagrinėkime kaip keisis kūdikių svoriai prie 0.5 kvantilio. Gauname tokį modelį:

$$Naujagimio\ svoris = -97.06 \cdot (\tau) + 0.48 \cdot (\tau) \cdot nėštumo\ laikotarpis - 3.43 \cdot (\tau) \cdot pirmas\ nėštumas(ne) + 1.26 \cdot (\tau) \cdot ūgis + 0.05 \cdot (\tau) \cdot svoris - 8.02 \cdot rūko(taip), kur\ \tau = 0.5.$$

Prie 0.5 kvantilio ilgėjant nėštumo laikotarpiui, naujagimio svoris padidėja 0.48 uncijomis. Jei moteriai tai buvo ne pirmas nėštumas, naujagimis svers 3.43 uncijomis mažiau. Jei moteris yra aukštesnė, kūdikis svers 1.26 uncijomis daugiau. Esant didesniam nėščiosios svoriui, naujagimio svoris padidės 0.05 uncijomis. Jei moteris rūko, tuomet naujagimio svoris bus mažesnis 8.02 uncijomis.

Taip pat pabandėme į modelį įtraukti sąveikų. Radome vieną sąveiką tarp nėštumo laikotarpio ir rūkymo, kuri atlikus pažingsninę regresiją išlieka reikšminga, o modelio AIC reikšmė sumažėja – 6952.59. Įtraukus sąveiką, prie 0.5 kvantilio gauname tokį modelį:

$$Naujagimio\ svoris = -82.63 \cdot (\tau) + 0.44 \cdot (\tau) \cdot nėštumo\ laikotarpis - 3.44 \cdot (\tau) \cdot pirmas\ nėštumas(ne) + 1.89 \cdot (\tau) \cdot ūgis + 0.06 \cdot (\tau) \cdot svoris - 52 \cdot (\tau) \cdot rūko(taip) + 0.16 \cdot (\tau) \cdot (nėštumo\ laikotarpis \cdot rūko(taip)), kur\ \tau = 0.5.$$

Prie 0.5 kvantilio ilgėjant nėštumo laikotarpiui, naujagimio svoris padidėja 0.44 uncijomis. Jei moteriai tai buvo ne pirmas nėštumas, naujagimis svers 3.44 uncijomis mažiau. Jei moteris yra aukštesnė, kūdikis svers 1.89 uncijomis daugiau. Esant didesniam nėščiosios svoriui, naujagimio svoris padidės 0.06 uncijomis. Jei moteris rūko, tuomet naujagimio svoris bus mažesnis 52 uncijomis. Įtraukta sąveika rodo, kad rūkančių moterų naujagimio svoris padidės per 0.16 uncijų esant ilgesniam nėštumo laikotarpiui, palyginti su moterimis, kurios nerūko.

Pastebėjome, kad modeliuojant skirtingo lygio kvantilius gauname skirtingus reikšmingų kovariančių rinkinius. Pavyzdžiui, prie 0.1 ir 0.9 kvantilio gaunamos tokios reikšmingos kovariantės (6 ir 7 lentelės). Matome, jog nėštumo laikotarpio, pirmojo nėštumo, ūgio ir rūkymo kovariantės trijuose modeliuose prie skirtingo lygmens kvantilių išlieka reikšmingiausiomis.

6 lentelė. Reikšmingos kovariantės prie 0.1 kvantilio

Kovariantės	**Nėštumo laikotarpis**	**Pirmas nėštumas (ne)**	**Amžius**	**Ūgis**	**Rūko (taip)**
Koeficientas	0.46	- 3.77	- 0.18	0.84	- 8.76

7 lentelė. Reikšmingos kovariantės prie 0.9 kvantilio

Kovariantės	**Nėštumo laikotarpis**	**Pirmas nėštumas (ne)**	**Ūgis**	**Svoris**	**Rūko (taip)**
Koeficientas	0.37	- 7.95	1.28	0.16	- 9.27

Taip pat iš testavimo aibės pateikiame 10 prognozuotų reikšmių naudojant modelį su sąveika ir be jos prie 0.5 kvantilio. Iš lentelės (8 lentelė) matome, jog abu modeliai prognozuoja panašiai.

8 lentelė. Prognozuotos reikšmės

Stebėtos reikšmės	Prognozuotos reikšmės	Prognozuotos reikšmės (su sąveika)
136	119.76	119.06
132	108.28	114.53
92	102.53	98.01
119	120.94	123.2
105	105	106.23
131	129.83	127.21
122	112.09	113.9
93	119.69	118.6
134	127.75	127.38
122	105.98	104.86

3. IŠVADOS

Iš pirminės duomenų analizės sužinojome, jog vidutinis nėštumo laikotarpis yra 279 dienos, moters amžius – 27 metai. 73 % moterų, tai buvo pirmasis nėštumas ir beveik 40 % nėščiųjų buvo rūkančios.

Duomenys tenkino kvantilių regresijos prielaidas – tiesiškumo sąlyga, išskirčių nebuvimas.

Atlikus pažingsninę regresiją nereikšminga tapo amžiaus kovariantė, todėl į modelį jos neįtraukėme. Į modelį įtraukta sąveika tarp nėštumo laikotarpio ir rūkumo sumažino jo AIC koeficientą (6952.59). Tačiau abu modeliai (prie 0.5 kvantilio), tiek su sąveika, tiek be jos, naujagimių svorį prognozuoja panašiai. Taigi, gavome tokį modelį:

$$Naujagimio\ svoris = -97.06 \cdot (\tau) + 0.48 \cdot (\tau) \cdot nėštumo\ laikotarpis - 3.43 \cdot (\tau) \cdot pirmas\ nėštumas(ne) + 1.26 \cdot (\tau) \cdot ūgis + 0.05 \cdot (\tau) \cdot svoris - 8.02 \cdot rūko(taip), kur\ \tau = 0.5.$$

Prie 0.5 kvantilio ilgėjant nėštumo laikotarpiui, naujagimio svoris padidėja 0.48 uncijomis. Jei moteriai tai buvo ne pirmas nėštumas, naujagimis svers 3.43 uncijomis mažiau. Jei moteris yra aukštesnė, kūdikis svers 1.26 uncijomis daugiau. Esant didesniam nėščiosios svoriui, naujagimio svoris padidės 0.05 uncijomis. Jei moteris rūko, tuomet naujagimio svoris bus mažesnis 8.02 uncijomis. Matome, jog didesnį naujagimio svorį nulemia ilgesnis nėštumo laikotarpis, didesnis ūgis bei svoris. Moters ne pirmas nėštumas ir rūkymas yra neigiamai susiję su naujagimio svoriu.