Univerzita Pardubice

Fakulta elektrotechniky a informatiky

Zpracování dat pro předmět NMAST

Bc. Lukáš Milar, Bc. Tomáš Prudký

Semestrální práce

2021

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Tato semestrální práce se zabývá analýzou vývoje epidemie nemoci Covid-19 v ČR. Za tímto účelem jsou srovnány přírůstky nových případů s našimi sousedy, efektivita testů při odhalování nových případů, úmrtnost nakažených, střední hodnota hospitalizovaných a vývoj střední hodnoty nových případů v čase. Dále je provedeno srovnání středních hodnot nových případů na milión s našimi sousedy a zkoumáno jaké rozdělení pravděpodobnosti data následují. Nakonec je pomocí regrese analyzováno podle jaké funkce se řídí přírůstek nových padientů na ICU v závislosti na nových hospitalizacích. Použitá data čerpají ze zdroje [1].

1 POPIS DAT

Data použitá v této práci se zabývají veličinami ohledně nemoci Covid-19 a pochází od společnosti Our World in Data. Tato data jsou denně aktualizována a obsahují například informace o očkování, testech, hospitalizacích, nových případech, nových úmrtích či reprodukčním čísle. Veškeré hodnoty jsou pozorovány napříč mnoha státy. Pro bližší popis těchto dat vizte zdroj [1].

Covid-19 (též COVID-19; z anglického spojení coronavirus disease 2019, což česky znamená koronavirové onemocnění 2019; výslovnost: [kovid devatenáct]; podle ICD-11 označené XN109) je vysoce infekční onemocnění, které je způsobeno koronavirem SARS-CoV-2. První případ byl identifikován v čínském Wu-chanu v prosinci 2019. Od té doby se virus rozšířil po celém světě, což způsobilo přetrvávající pandemii.

Příznaky nemoci covid-19 jsou různé, od bezpříznakového stavu až po závažné onemocnění, ale často zahrnují horečku, kašel, únavu, dýchací potíže a ztrátu čichu a chuti. Příznaky začínají jeden až čtrnáct dní po vystavení viru. U přibližně jednoho z pěti infikovaných jedinců se neobjeví žádné příznaky. Zatímco většina lidí má mírné příznaky, u některých lidí se vyvine syndrom akutní dechové tísně. Tento syndrom může být přivoděn cytokinovými bouřemi, víceorgánovým selháním, septickým šokem a krevními sraženinami. Bylo pozorováno dlouhodobější poškození orgánů (zejména plic a srdce). Existuje obava z významného počtu pacientů, kteří se zotavili z akutní fáze onemocnění, ale nadále pocit'ují řadu následků – známých jako dlouhodobý covid-19 – i několik měsíců poté. Mezi tyto účinky patří silná únava, ztráta paměti a další kognitivní problémy, slabá horečka, svalová slabost a dušnost.

Virus, který způsobuje covid-19, se šíří hlavně vzdušným přenosem, když je infikovaná osoba v blízkém kontaktu s jinou osobou. Malé kapičky a aerosoly obsahující virus se mohou šířit z nosu a úst infikované osoby při dýchání, kašlání, kýchání, zpěvu nebo mluvení. Ostatní lidé se mohou nakazit, pokud se virus dostane do jejich úst, nosu nebo očí. Virus se může šířit také kontaminovaným povrchem, i když to není považováno za hlavní cestu přenosu. Přesná cesta přenosu je zřídkakdy přesvědčivě prokázána, ale k infekci dochází hlavně tehdy, když jsou lidé dostatečně blízko sebe. Virus se může šířit až dva dny předtím, než infikované osoby projeví příznaky, a od jedinců, kteří nikdy nepocit'ují příznaky. Lidé zůstávají infekční po dobu až deseti dnů při středně závažných případech a dva týdny ve

vážných případech. Virus se šíří snadněji ve vnitřních prostorách a v davu. Pro diagnózu onemocnění byly vyvinuty různé testovací metody. Standardní diagnostickou metodou je reverzní transkripční polymerázová řetězová reakce v reálném čase (PCR test) výtěrem z nosohltanu.

Preventivní opatření zahrnují fyzický či společenský odstup, umístění ohrožených osob do karantény, větrání vnitřních prostor, zakrývání úst a nosu při kašli a kýchání, mytí rukou a udržování neumytých rukou pryč od obličeje. Aby se minimalizovalo riziko přenosu, bylo na veřejnosti doporučeno použití roušek, obličejových masek nebo jiného zakrytí dýchacích cest. Bylo vyvinuto několik vakcín proti covidu-19, načež většina států světa zahájila očkovací kampaně a samotné očkování, jehož rozsah je ovšem závislý na přístupnosti dostatečného množství vakcín.

Ačkoli probíhají práce na vývoji léků, které zpomalují a zastavují virus, primární léčba je v současnosti symptomatická. Zahrnuje léčbu příznaků, podpůrnou péči, izolaci a některá experimentální opatření.

Vzhledem k velkému množství dat jsou v této práci použity zpravidla údaje pro Českou republiku, ze kterých je dále využit užší výčet dostupných veličin.

2 POPISNÁ STATISTIKA

V tabulkách níže jsou zobrazeny hodnoty popisné statistiky pro veličiny nových případů, 7denního klouzavého průměru nových případů, nových případů na milion, 7denního klouzavého průměru nových případů na milion, hospitalizovaných pacientů a hospitalizovaných pacientů na milion v České republice. Hodnoty 7denního klouzavého průměru lépe zachycují tyto veličiny v rámci dlouhodobých trendů, jelikož je eliminováno zkreslení v podobě menšího počtu uskutečněných testů například během víkendů.

	Nové případy	7denní klouzavý průměr nových případů
průměr	2940.58	2936.89
modus	75.00	57.57
medián	416.00	422.29
max	17773.00	12954.86
min	-2214.00	2.71
šikmost	1.55	1.16
špičatost	1.38	-0.07
odchylka	4277.10	3876.20
variance	18293577.15	15024928.55

Tabulka 1: Části popisné statistiky aplikované na data nových případů a jejich 7denního klouzavého průměru v ČR od 7. 3. 2020

	Nové případy na milión	7denní klouzavý průměr nových případů na milión
průměr	274.19	273.85
modus	6.99	5.37
medián	38.79	39.38
max	1657.22	1207.96
min	-206.44	0.25
šikmost	1.55	1.16
špičatost	1.38	-0.07
odchylka	398.81	361.43
variance	159052.40	130633.34

Tabulka 2: Části popisné statistiky aplikované na data nových případů na milión a jejich 7denního klouzavého průměru v ČR od 7. 3. 2020

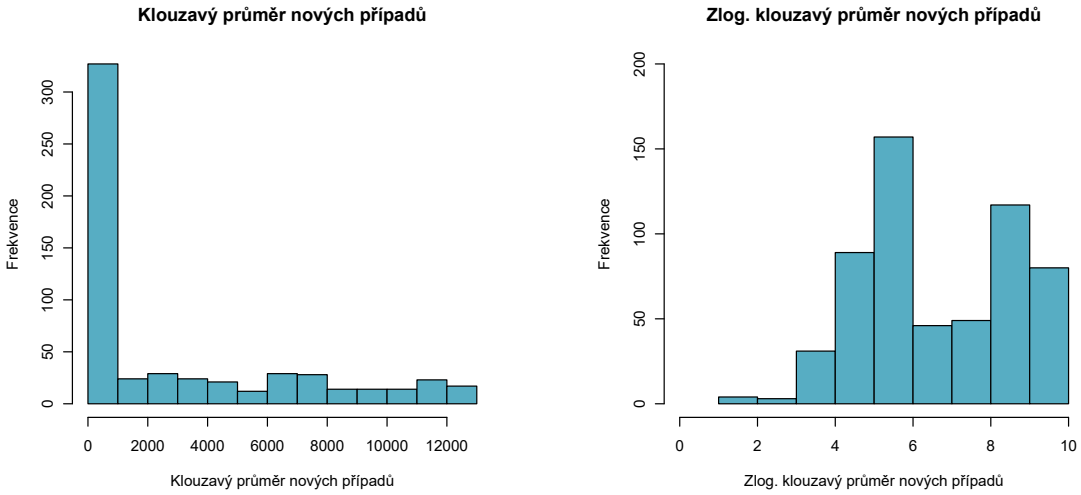
	Hospitalizovaní pacienti	Hospitalizovaní pacienti na milión
průměr	2370.49	221.03
modus	69.00	6.43
medián	339.00	31.61
max	9509.00	886.66
min	0.00	0.00
šikmost	0.86	0.86
špičatost	-0.86	-0.86
odchylka	2973.01	277.22
variance	8838793.84	76848.36

Tabulka 3: Části popisné statistiky aplikované na data nových hospitalizací a nových hospitalizací na milión v ČR od 7. 3. 2020

3 ZÁKLADNÍ GRAFY

3.1 Histogram

Následující histogram zobrazuje četnost hodnot klouzavého průměru nových případů v ČR od 7. 3. 2020. Vzhledem k očividnému zešikmení dat vlevo byla pro lepší přehlednost data zlogaritmována.

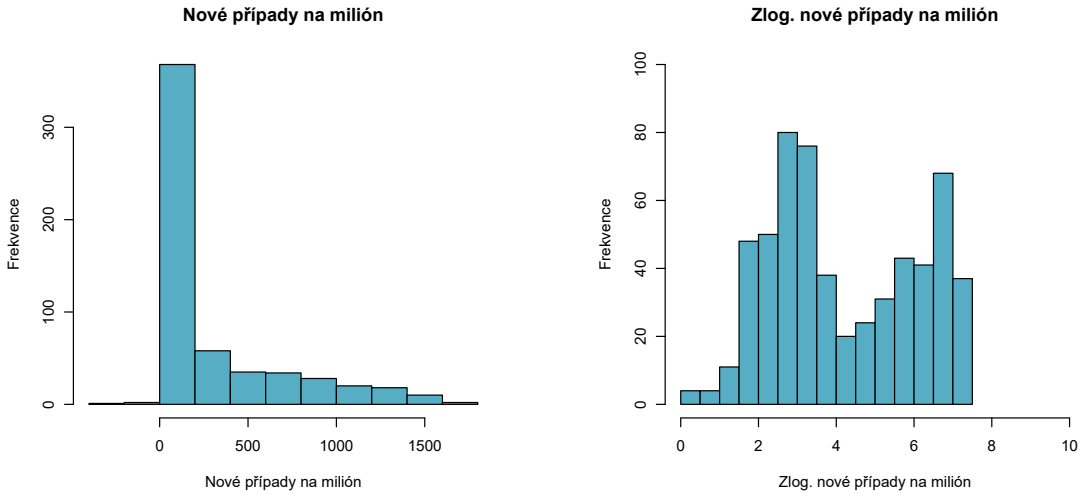


(a) Klouzavý průměr nových případů

(b) Logaritmická transformace

Obrázek 1: Klouzavý průměr nových případů v ČR od 7. 3. 2020

Následující histogram zobrazuje četnost hodnot nových případů na milión obyvatel v ČR od 7. 3. 2020. Vzhledem k očividnému zešikmení dat vlevo byla pro lepší přehlednost data opět zlogaritmována.

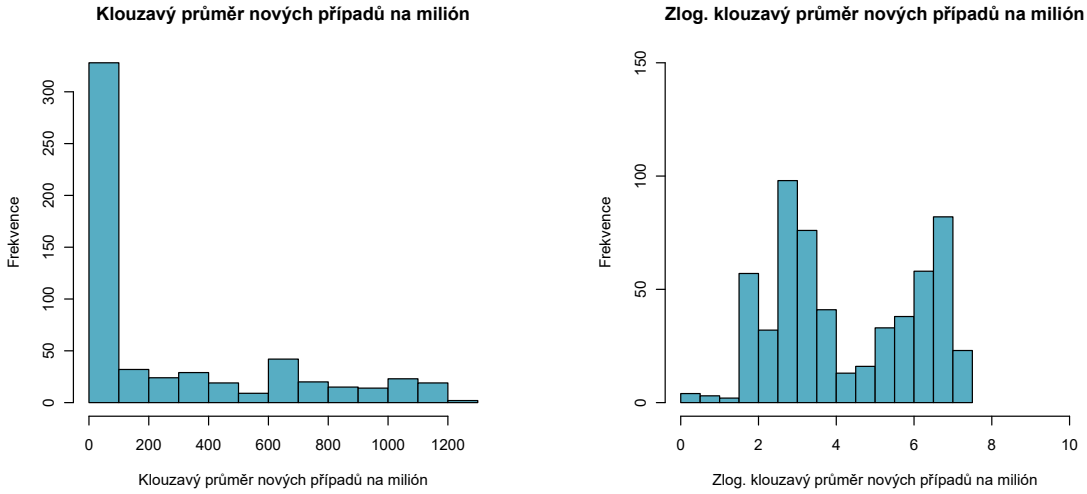


(a) Nové případy na milión

(b) Logaritmická transformace

Obrázek 2: Nové případy na milión v ČR od 7. 3. 2020

Následující histogram zobrazuje četnost hodnot 7denního klouzavého průměru nových případů na milión obyvatel v ČR od 7. 3. 2020. Vzhledem k očividnému zešikmení dat vlevo byla pro lepší přehlednost data opět zlogaritmována.

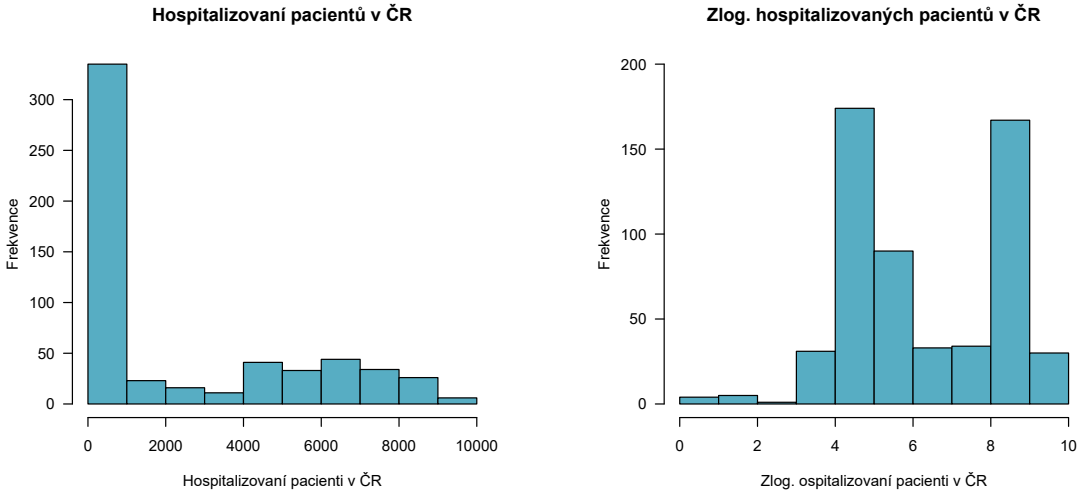


(a) Klouzavý průměr nových případů na milión

(b) Logaritmická transformace

Obrázek 3: Klouzavý průměr nových případů na milión v ČR od 7. 3. 2020

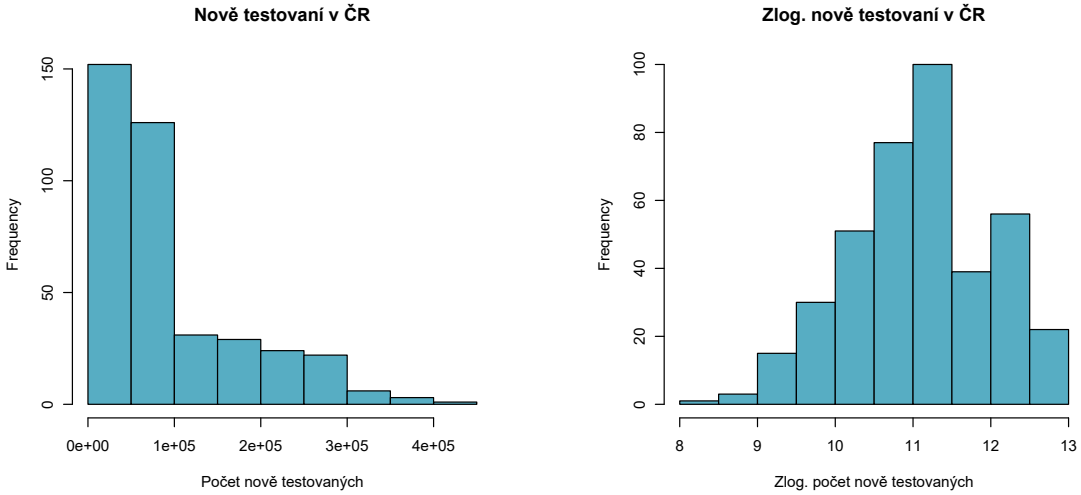
Následující histogram zobrazuje četnost hodnot hospitalizovaných pacientů v ČR od 7. 3. 2020. Vzhledem k očividnému zešikmení dat vlevo byla pro lepší přehlednost data opět zlogaritmována.



(a) Hospitalizovaní pacienti (b) Logaritmická transformace

Obrázek 4: Hospitalizovaní pacienti v ČR od 7. 3. 2020

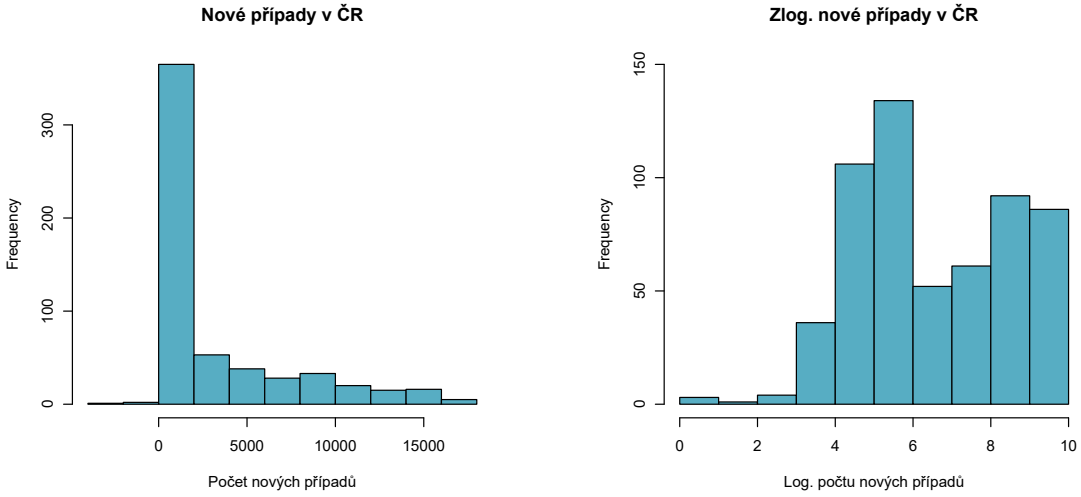
Následující histogram zobrazuje četnost hodnot nově testovaných v ČR od 7. 3. 2020. Vzhledem k očividnému zešikmení dat vlevo byla pro lepší přehlednost data opět zlogaritmována.



(a) Nově testovaní (b) Logaritmická transformace

Obrázek 5: Nově testovaní v ČR od 7. 3. 2020

Následující histogram zobrazuje četnost hodnot nových případů v ČR od 7. 3. 2020. Vzhledem k očividnému zešikmení dat vlevo byla pro lepší přehlednost data opět zlogaritmována.

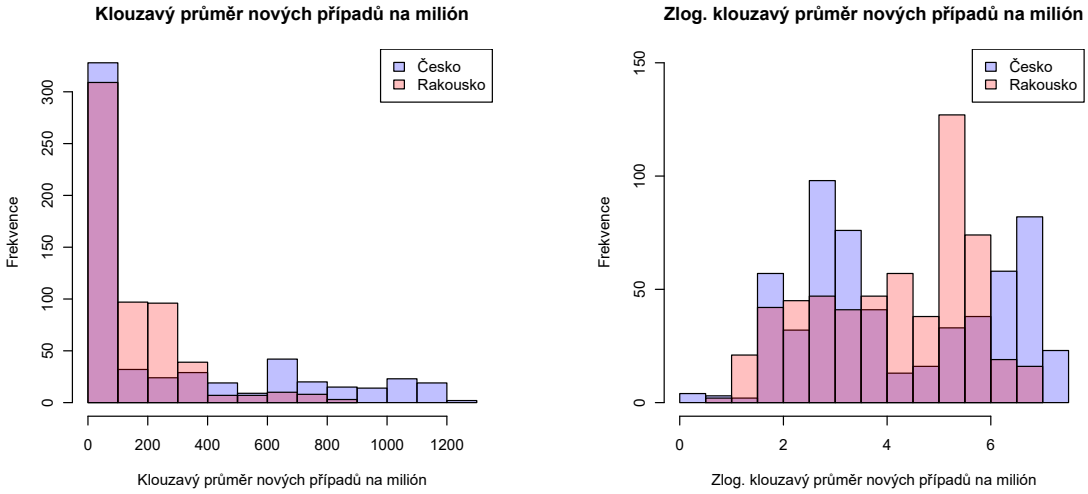


(a) Nové případy

(b) Logaritmická transformace

Obrázek 6: Nové případy v ČR od 7. 3. 2020

Následující histogram zobrazuje srovnání četnosti hodnot klouzavého průměru nových případů v ČR a Rakousku od 7. 3. 2020. Vzhledem k očividnému zešikmení dat vlevo byla pro lepší přehlednost data opět zlogaritmována.

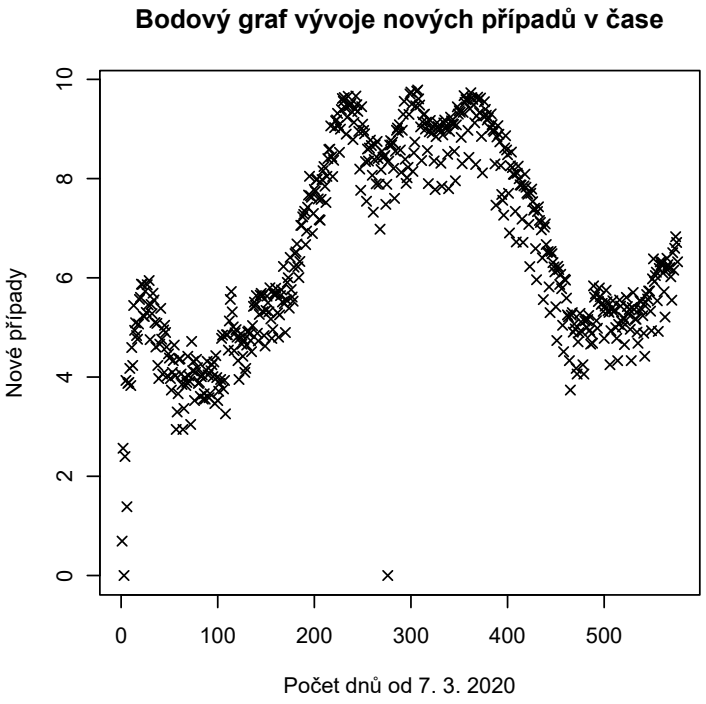


(a) Nové případy na milión pro ČR a Rakousko

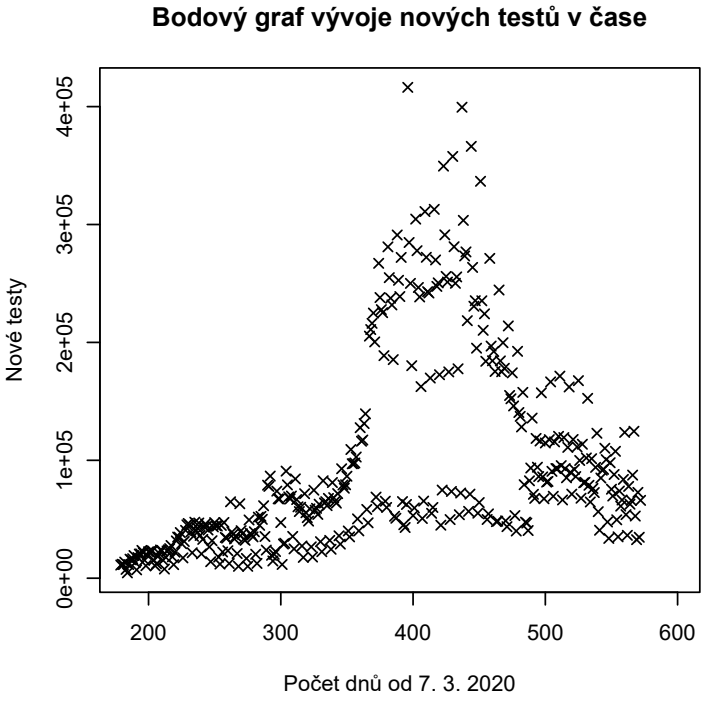
(b) Logaritmická transformace

Obrázek 7: Klouzavý průměr nových případů na milión pro Česko a Rakousko od 7. 3. 2020

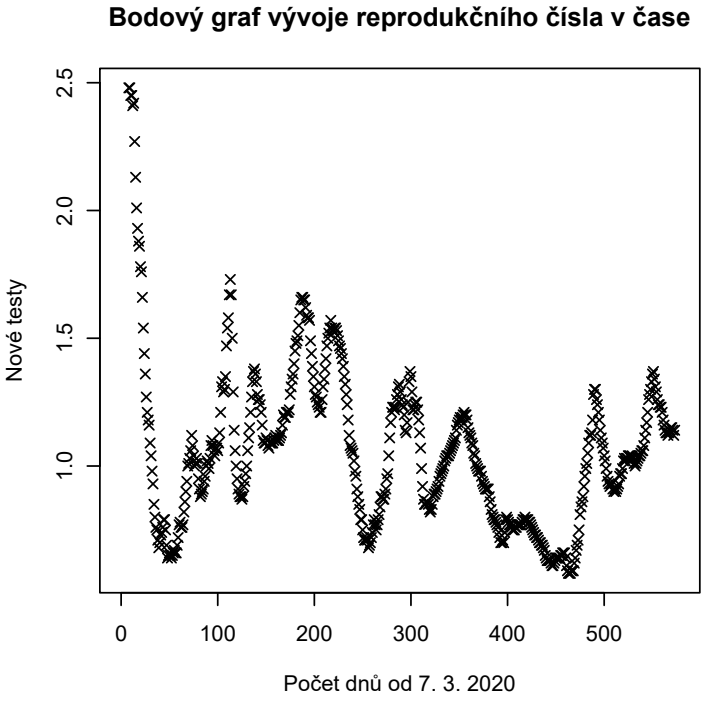
3.2 Bodový graf



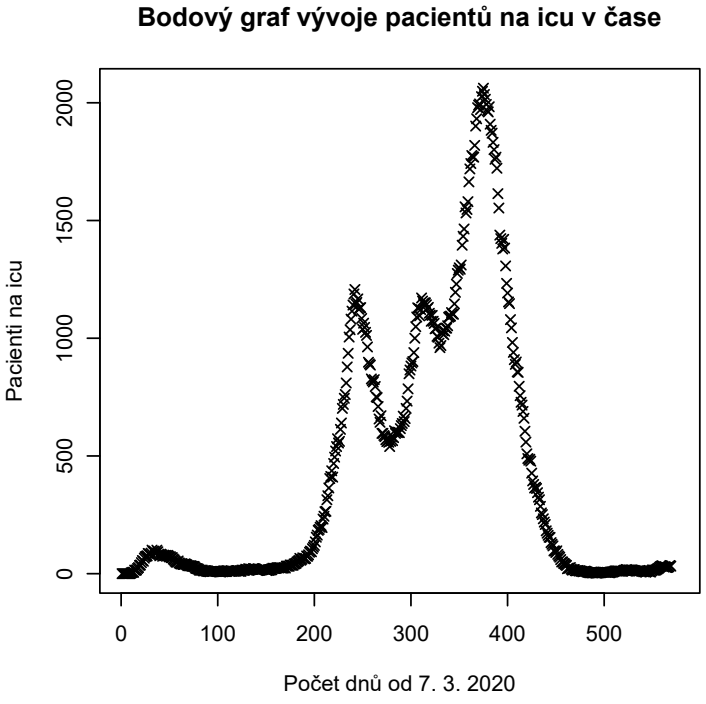
Obrázek 8: Bodový graf zlogaritmovaných nových případů



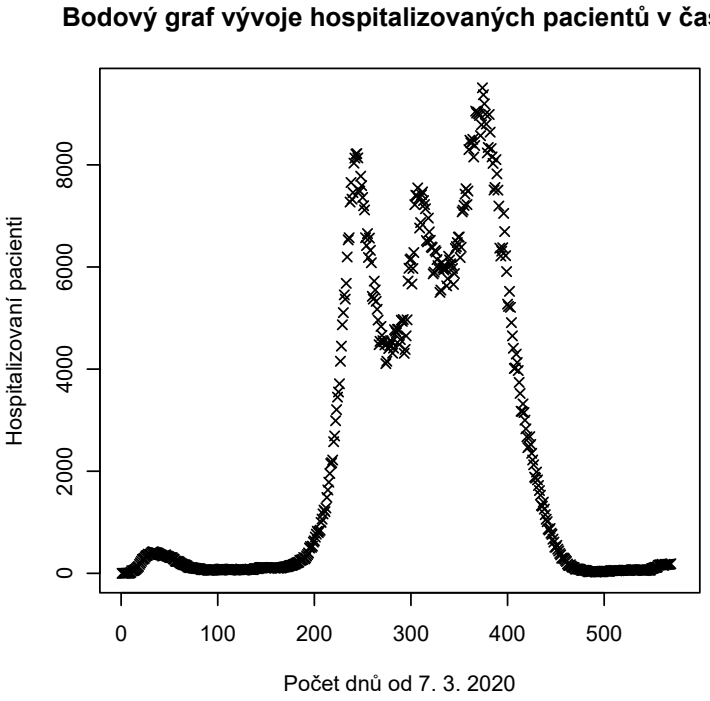
Obrázek 9: Bodový graf nových testů



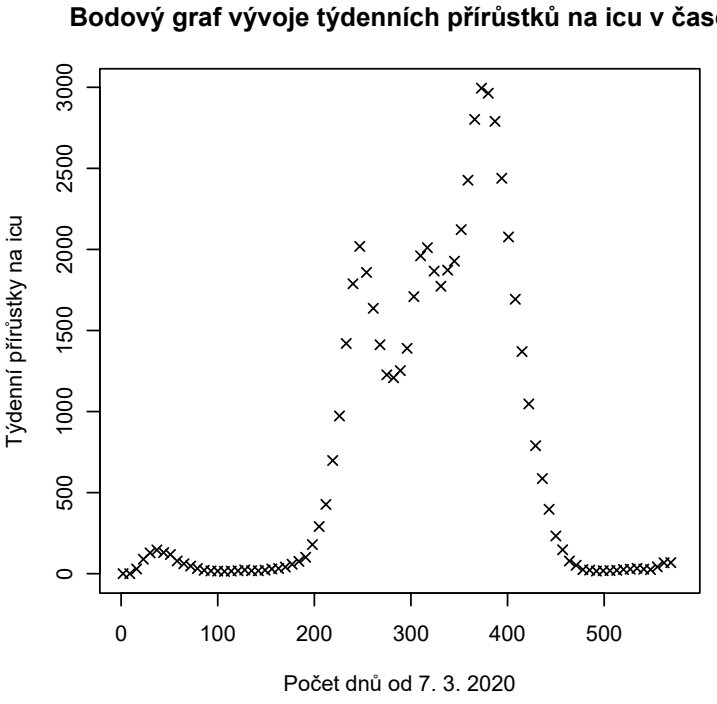
Obrázek 10: Bodový graf reprodukčního čísla



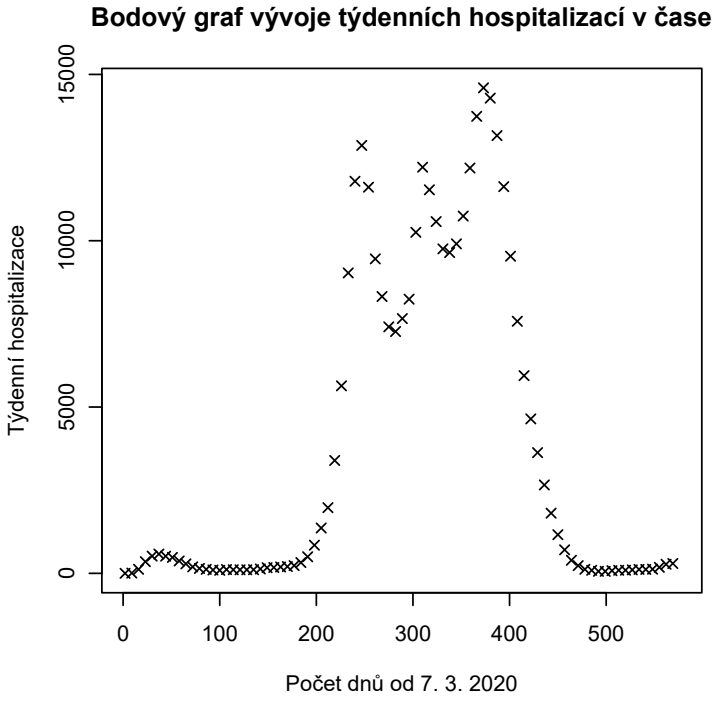
Obrázek 11: Bodový graf pacientů na icu



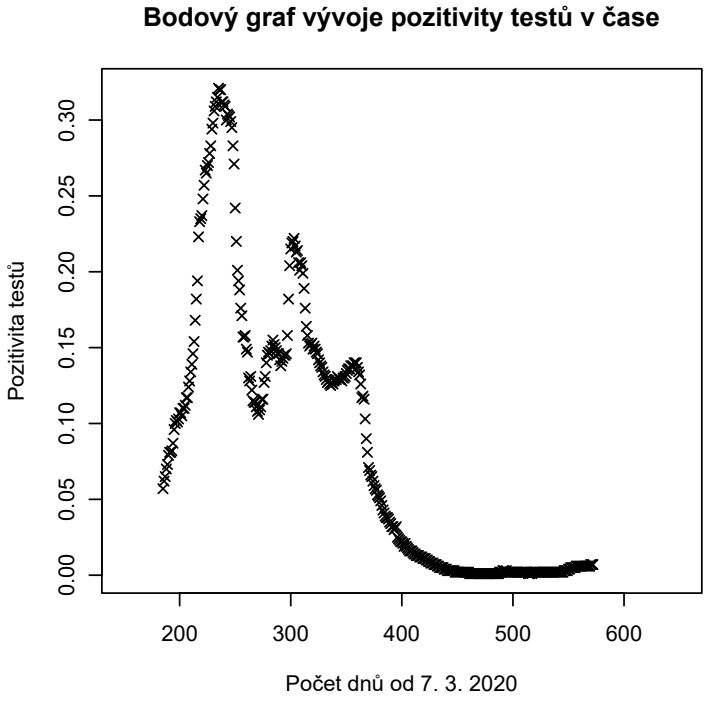
Obrázek 12: Bodový graf hospitalizovaných pacientů



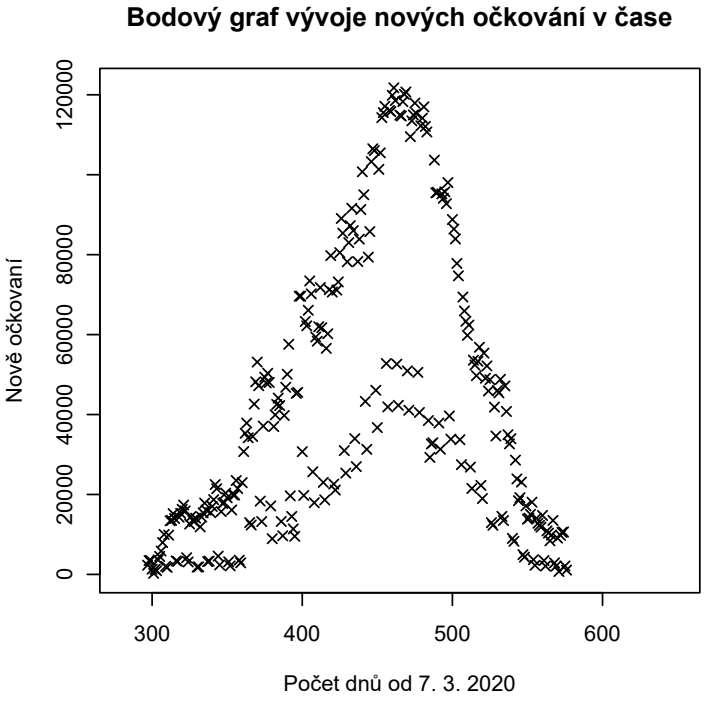
Obrázek 13: Bodový graf týdenních přírůstků na icu



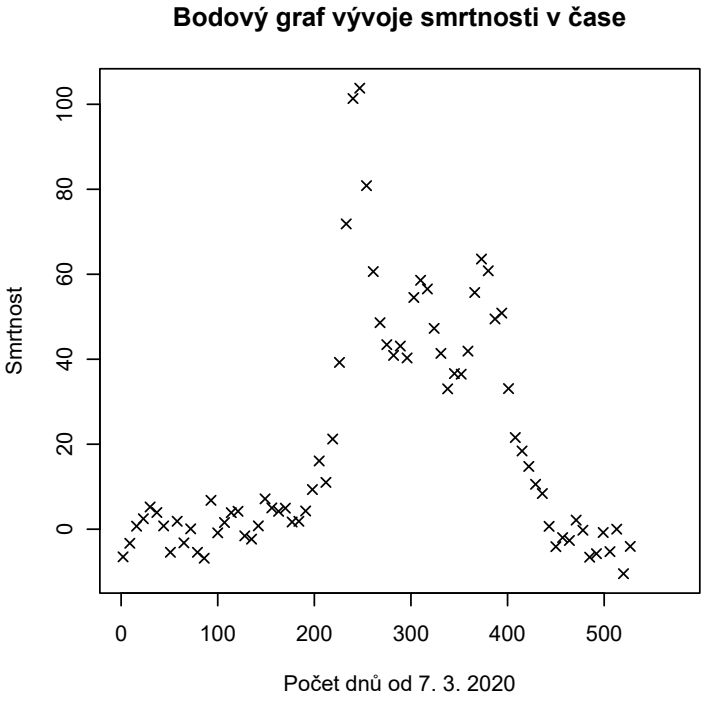
Obrázek 14: Bodový graf týdenních hospitalizací



Obrázek 15: Bodový graf pozitivity testů

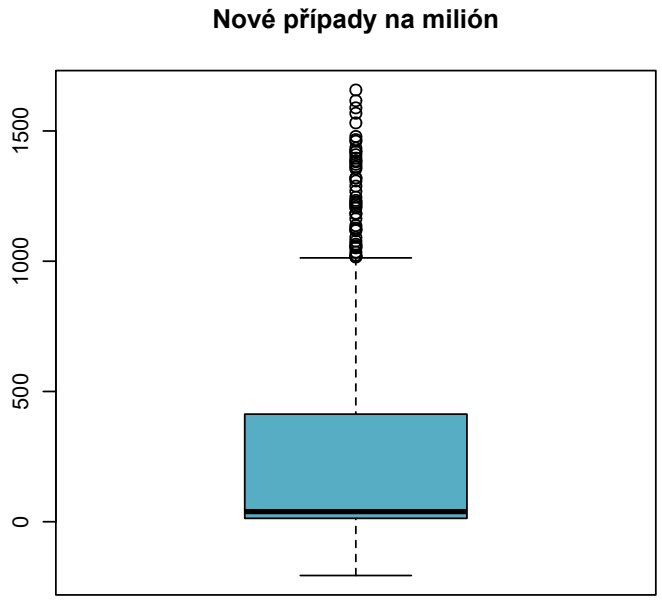


Obrázek 16: Bodový graf nových očkování

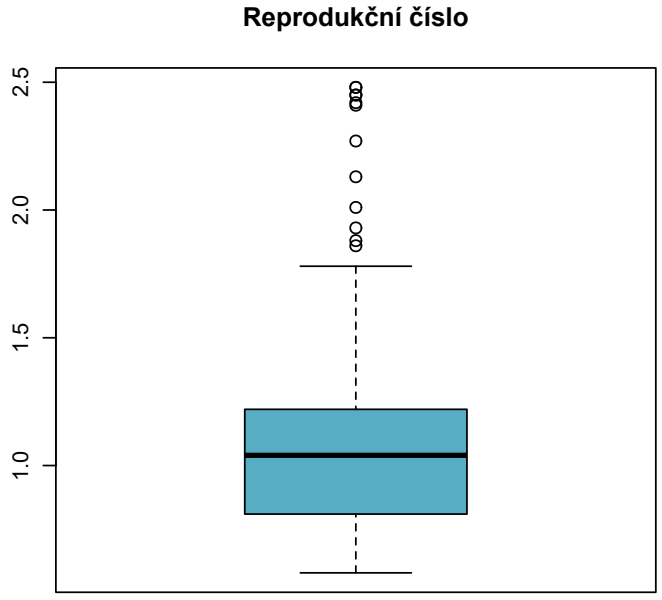


Obrázek 17: Bodový graf smrtnosti

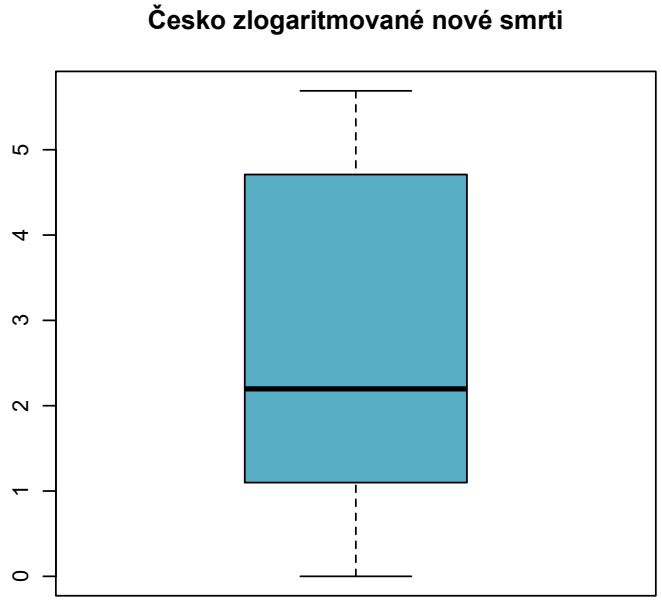
3.3 Boxplot



Obrázek 18: Boxplot graf pro nové případy na milión

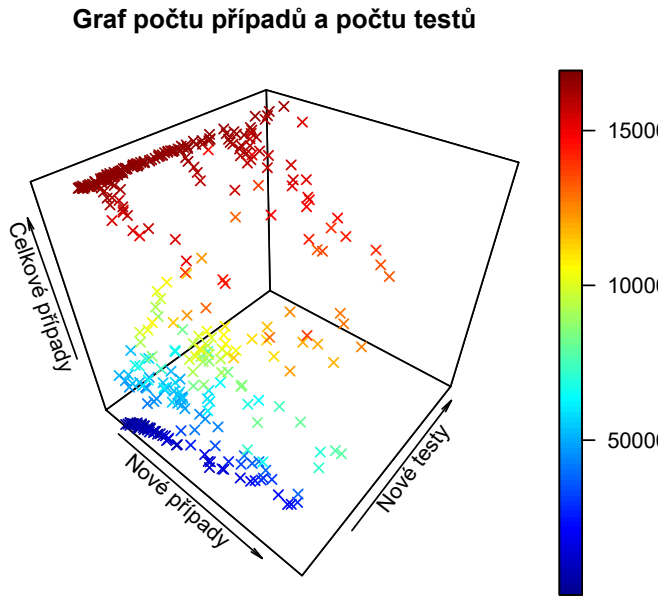


Obrázek 19: Boxplot graf pro reprodukční číslo

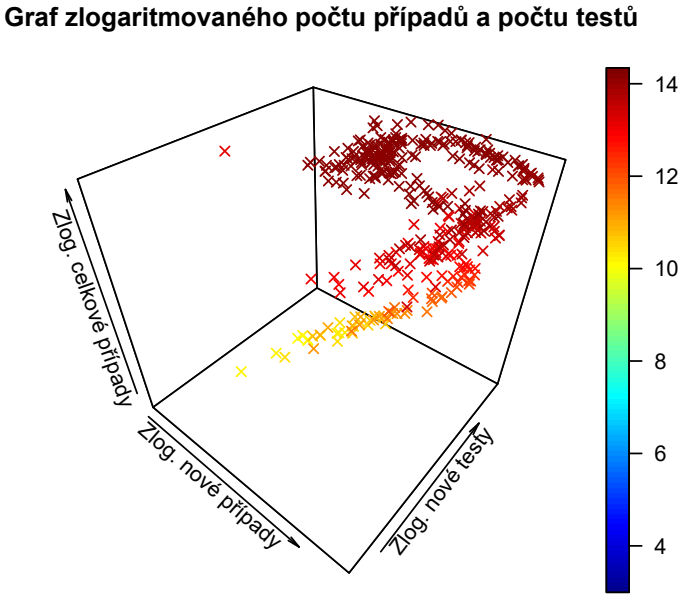


Obrázek 20: Boxplot graf pro zlogaritmované nové smrti

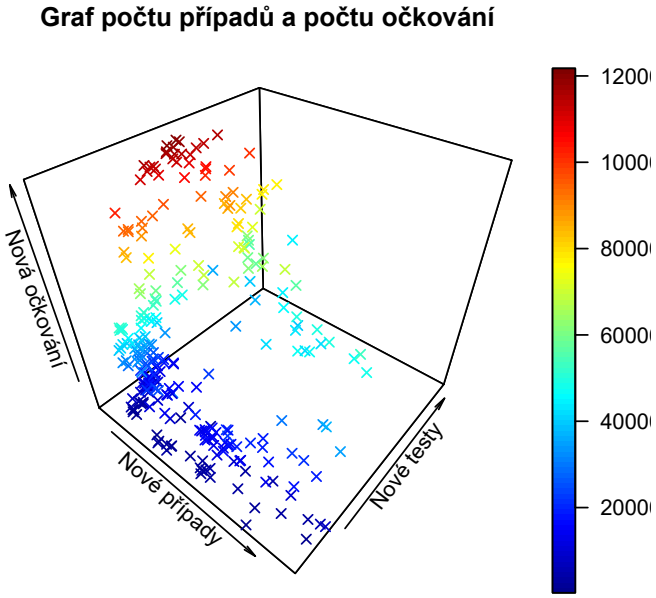
3.4 3D graf



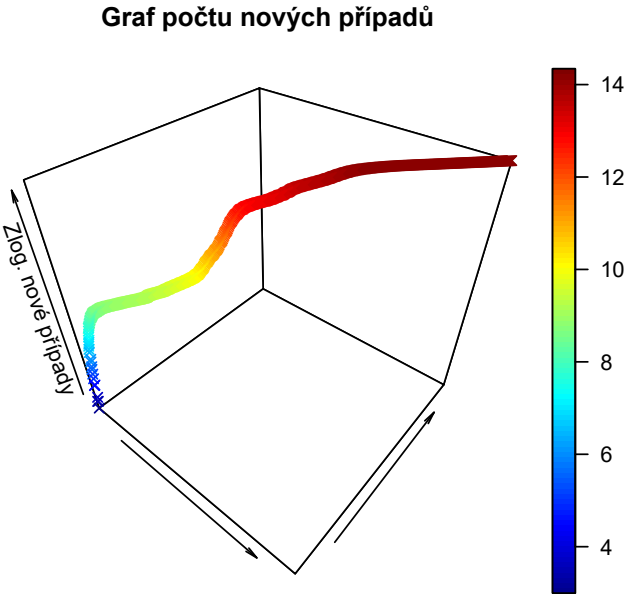
Obrázek 21: 3D graf počtu případů a počtu testů



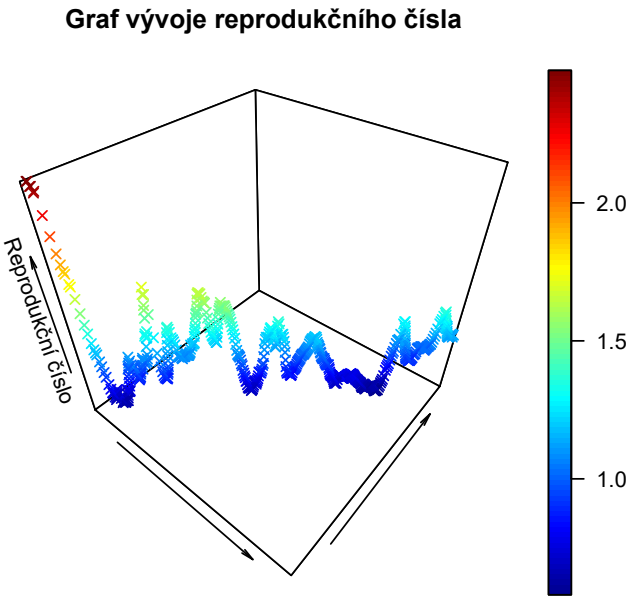
Obrázek 22: 3D graf zlogaritmovaných počtu případů a počtu testů



Obrázek 23: 3D graf počtu případů a počtu nových očkování

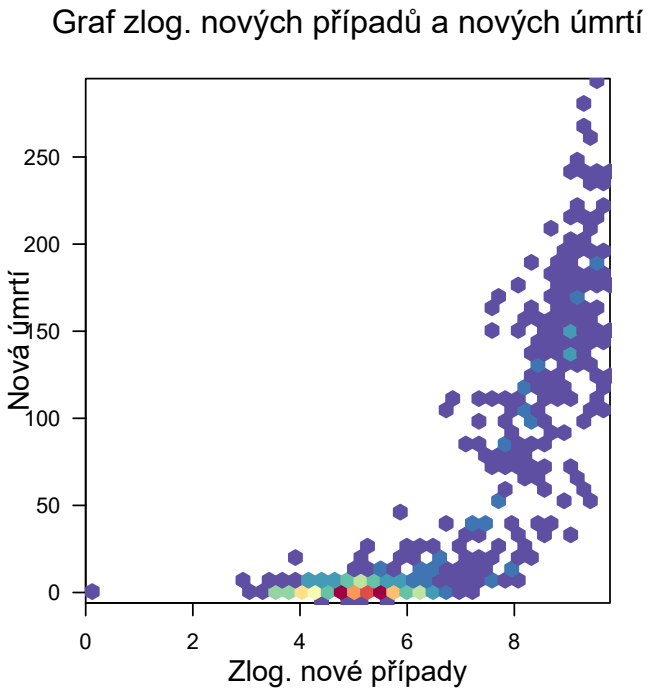


Obrázek 24: 3D graf počtu nových případů



Obrázek 25: 3D graf reprodukčního čísla

3.5 Hexbin



Obrázek 26: Hexbin graf nových zlog. nových případů a nových úmrtí

3.6 Chernoff faces

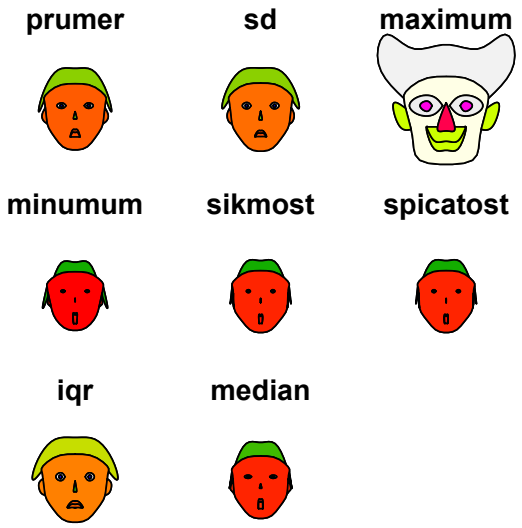
Pomocí chernoff faces jsou vyobrazeny obličeje na základě hodnot pro Česko. Hodnoty, které jsou použity jsou new_cases, new_deaths, new_tests, $new_vaccinations$, $icu_patiens$. Každý obličej reprezentuje jednu funkci, která je aplikována na zvolené hodnoty. Z grafu je zřejmé zešikmení dat vzhledem k velkým rozdílům mezi hodnotami průměru a mediánu zobrazovaných veličin.

	new_cases	new_deaths	new_tests	new_vaccinations	icu_patients
prumer	2940.58	54.33	94464.09	42220.25	424.51
sd	4277.10	72.32	83099.10	35340.13	558.31
maximum	17773.00	295.00	416333.00	121742.00	2062.00
minumum	-2214.00	-6.00	4537.00	266.00	0.00
sikmost	1.55	1.11	1.39	0.79	1.23
spicatost	1.38	-0.06	1.22	-0.57	0.43
iqr	4275.75	107.00	82303.50	49771.00	799.00
median	416.00	7.00	65211.00	33315.00	74.00

Tabulka 4: Hodnoty dat znázorněných pomocí Chernoffových obličejů

```
effect of variables:
 modified item       Var
 "height of face   " "new_cases"
 "width of face    " "new_deaths"
 "structure of face" "new_tests"
 "height of mouth  " "new_vaccinations"
 "width of mouth   " "icu_patients"
 "smiling          " "new_cases"
 "height of eyes   " "new_deaths"
 "width of eyes    " "new_tests"
 "height of hair   " "new_vaccinations"
 "width of hair   "  "icu_patients"
 "style of hair   "  "new_cases"
 "height of nose  "  "new_deaths"
 "width of nose   "  "new_tests"
 "width of ear    "  "new_vaccinations"
 "height of ear   "  "icu_patients"
```

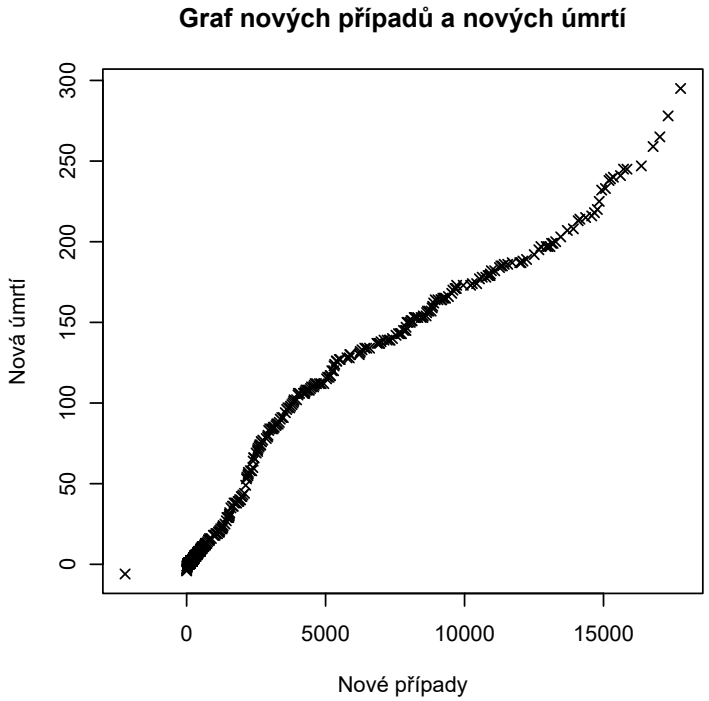
Obrázek 27: Legenda Chernoff faces grafu tabulky popisné statistiky



Obrázek 28: Chernoff faces graf tabulky popisné statistiky

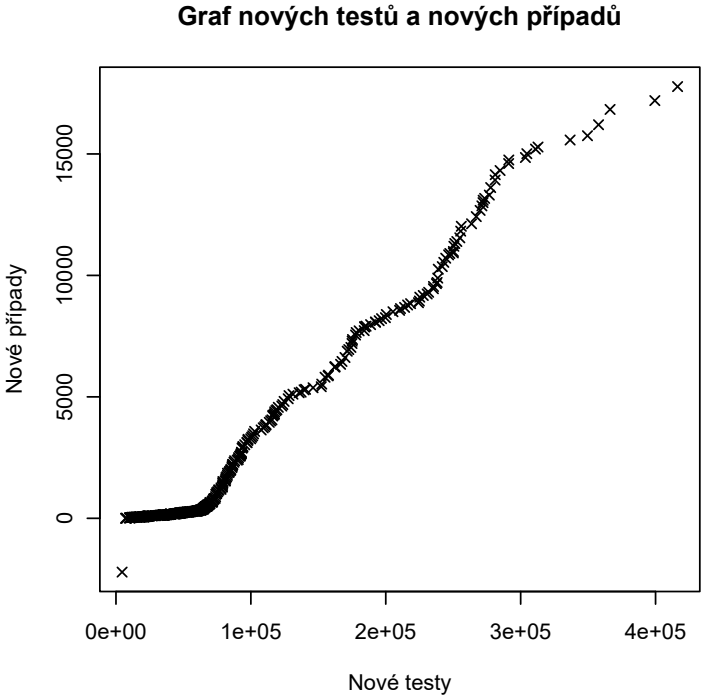
3.7 QQPlot

Na základě následujícího QQPlot grafu je možné dojít k závěru, že počty nových úmrtí a počty nových případů v ČR se řídí dle podobného rozdělení pravděpodobnosti.



Obrázek 29: QQPlot graf nových případů a nových úmrtí

Na základě následujícího QQPlot grafu je možné dojít k závěru, že počty nových případů a počty nových testů v ČR se řídí dle podobného rozdělení pravděpodobnosti.



Obrázek 30: QQPlot graf nových testů a nových případů

4 TESTOVÁNÍ STATISTICKÝCH HYPOTÉZ

4.1 Jednovýběrový Studentův test vůči střední hodnotě

Následující test testuje zda se střední hodnota nových případů v ČR rovná hodnotě 3300 s hladinou významnosti $\alpha = 0.05$. Testová statistika nabývá hodnoty -2.0168 při 575 stupních volnosti. Vzhledem k tomu, že hodnota p-value je nižší než hladina významnosti, tuto hypotézu zamítáme ve prospěch hypotézy alternativní, tudíž že se střední hodnota nových případů v ČR nerovná hodnotě 3300.

```
        One Sample t-test


data:  new_cases_czechia
t = -2.0168, df = 575, p-value = 0.04418
alternative hypothesis: true mean is not equal to 3300
95 percent confidence interval:
 2590.550 3290.603
sample estimates:
mean of x
 2940.576
```

Následující test testuje zda se střední hodnota 7denního klouzavého průměru nových případů v ČR rovná hodnotě 3300 s hladinou významnosti $\alpha = 0.05$. Testová statistika nabývá hodnoty -2.2482 při 575 stupních volnosti. Vzhledem k tomu, že hodnota p-value je nižší než hladina významnosti, tuto hypotézu zamítáme ve prospěch hypotézy alternativní, tudíž že se střední hodnota nových případů v ČR nerovná hodnotě 3300.

```
        One Sample t-test

data:  new_cases_smoothed_czechia
t = -2.2482, df = 575, p-value = 0.02494
alternative hypothesis: true mean is not equal to 3300
95 percent confidence interval:
 2619.672 3254.109
sample estimates:
mean of x
  2936.89
```

Následující test testuje zda se střední hodnota nových případů na milión v ČR rovná hodnotě 300 s hladinou významnosti $\alpha = 0.05$. Testová statistika nabývá hodnoty -1.5531 při 575 stupních volnosti. Vzhledem k tomu, že hodnota p-value je vyšší než hladina významnosti, tuto hypotézu nemůžeme zamítnout ve prospěch hypotézy alternativní.

```
        One Sample t-test

data:  new_cases_per_million_czechia
t = -1.5531, df = 575, p-value = 0.1209
alternative hypothesis: true mean is not equal to 300
95 percent confidence interval:
 241.5531 306.8289
sample estimates:
mean of x
  274.191
```

Následující test testuje zda se střední hodnota 7denního klouzavého průměru nových případů na milión v ČR rovná hodnotě 300 s hladinou významnosti $\alpha = 0.05$. Testová statistika nabývá hodnoty -1.7366 při 575 stupních volnosti. Vzhledem k tomu, že hodnota p-value je vyšší než hladina významnosti, tuto hypotézu nemůžeme zamítnout ve prospěch hypotézy alternativní.

```
        One Sample t-test

data:  new_cases_smoothed_per_million_czechia
t = -1.7366, df = 575, p-value = 0.08299
alternative hypothesis: true mean is not equal to 300
95 percent confidence interval:
 244.2687 303.4260
sample estimates:
mean of x
 273.8474
```

Následující test testuje zda se střední hodnota nových hospitalizovaných pacientů v ČR rovná hodnotě 2000 s hladinou významnosti $\alpha = 0.05$. Testová statistika nabývá hodnoty 2.9726 při 568 stupních volnosti. Vzhledem k tomu, že hodnota p-value je nižší než hladina významnosti, tuto hypotézu zamítáme ve prospěch hypotézy alternativní, tudíž že se střední hodnota nových hospitalizovaných pacientů v ČR nerovná hodnotě 2000.

```
        One Sample t-test

data:  hosp_patients_czechia
t = 2.9726, df = 568, p-value = 0.003078
alternative hypothesis: true mean is not equal to 2000
95 percent confidence interval:
 2125.690 2615.294
sample estimates:
mean of x
 2370.492
```

Následující test testuje zda se střední hodnota nových hospitalizovaných pacientů na milión v ČR rovná hodnotě 200 s hladinou významnosti $\alpha = 0.05$. Testová statistika nabývá hodnoty 1.8099 při 568 stupních volnosti. Vzhledem k tomu, že hodnota p-value je vyšší než hladina významnosti, tuto hypotézu nemůžeme zamítnout ve prospěch hypotézy alternativní.

```
        One Sample t-test

data:  hosp_patients_per_million_czechia
t = 1.8099, df = 568, p-value = 0.07083
alternative hypothesis: true mean is not equal to 200
95 percent confidence interval:
 198.2078 243.8604
sample estimates:
mean of x
 221.0341
```

4.2 Dvouvýběrový Studentův test

Následující dvouvýběrový t-test testuje hypotézu, že střední hodnota nových případů v první části dat z ČR je rovna střední hodnotě v druhé části. Vzhledem ke skutečnosti, že p-value je menší než hladina významnosti ($\alpha = 0{,}05$), zamítáme tuto hypotézu ve prospěch alternativní. Při této hladině významnosti tudíž můžeme tvrdit, že střední hodnota nových případů v první části dat z ČR se nerovná střední hodnotě z druhé části.

```
        Welch Two Sample t-test

data:  new_cases_czechia_p1 and new_cases_czechia_p2
t = -4.518, df = 537.03, p-value = 7.683e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2272.4359  -895.1752
sample estimates:
mean of x mean of y
 2148.674  3732.479
```

Následující dvouvýběrový t-test testuje hypotézu, že střední hodnota nových případů v ČR je rovna střední hodnotě nových případů v Německu. Testová statistika nabývá hodnoty -11,133 při 843,1 stupních volnosti. Vzhledem ke skutečnosti, že p-value je menší než hladina významnosti ($\alpha = 0{,}05$), zamítáme tuto hypotézu ve prospěch alternativní. Při této hladině významnosti tudíž můžeme tvrdit, že střední hodnota nových případů v Německu se nerovná střední hodnotě nových případů v ČR.

```
        Welch Two Sample t-test

data:  new_cases_czechia and new_cases_germany
t = -11.133, df = 843.1, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5240.394 -3669.509
sample estimates:
mean of x mean of y
 2940.576  7395.528
```

Následující dvouvýběrový t-test testuje hypotézu, že střední hodnota nových případů na milión v ČR je rovna střední hodnotě nových případů na milión na Slovensku. Vzhledem ke skutečnosti, že p-value je menší než hladina významnosti ($\alpha = 0{,}05$), zamítáme tuto hypotézu ve prospěch alternativní. Při této hladině významnosti tudíž můžeme tvrdit, že střední hodnota nových případů na milión na Slovensku se nerovná střední hodnotě nových případů na milión v ČR.

```
        Welch Two Sample t-test

data:  new_cases_per_million_czechia and new_cases_per_million_slovakia
t = 7.7283, df = 817.84, p-value = 3.194e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 105.8863 177.9857
sample estimates:
mean of x mean of y
  274.191   132.255
```

Následující dvouvýběrový t-test testuje hypotézu, že střední hodnota nových případů na milión v ČR je rovna střední hodnotě nových případů na milión v Německu. Vzhledem ke skutečnosti, že p-value je menší než hladina významnosti ($\alpha = 0{,}05$), zamítáme tuto hypotézu ve prospěch alternativní. Při této hladině významnosti tudíž můžeme tvrdit, že střední hodnota nových případů na milión v Německu se nerovná střední hodnotě nových případů na milión v ČR.

```
        Two Sample t-test

data:  new_cases_per_million_czechia and new_cases_per_million_germany
t = 10.844, df = 1150, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 152.3817 219.7075
sample estimates:
mean of x mean of y
274.19103  88.14644
```

Následující dvouvýběrový t-test testuje hypotézu, že střední hodnota nových případů na milión v ČR je rovna střední hodnotě nových případů na milión v Polsku. Vzhledem ke skutečnosti, že p-value je menší než hladina významnosti ($\alpha = 0{,}05$), zamítáme tuto hypotézu ve prospěch alternativní. Při této hladině významnosti tudíž můžeme tvrdit, že střední hodnota nových případů na milión v Polsku se nerovná střední hodnotě nových případů na milión v ČR.

```
        Two Sample t-test

data:  new_cases_per_million_czechia and new_cases_per_million_poland
t = 7.5404, df = 1150, p-value = 9.477e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 103.9326 177.0431
sample estimates:
mean of x mean of y
 274.1910  133.7032
```

Následující dvouvýběrový t-test testuje hypotézu, že střední hodnota nových případů na milión v ČR je rovna střední hodnotě nových případů na milión v Rakousku. Vzhledem ke skutečnosti, že p-value je menší než hladina významnosti ($\alpha = 0{,}05$), zamítáme tuto hypotézu ve prospěch alternativní. Při této hladině významnosti tudíž můžeme tvrdit, že střední hodnota nových případů na milión v Rakousku se nerovná střední hodnotě nových případů na milión v ČR.

```
        Two Sample t-test

data:  new_cases_per_million_czechia and new_cases_per_million_austria
t = 7.2242, df = 1150, p-value = 9.157e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  95.01377 165.86690
sample estimates:
mean of x mean of y
 274.1910  143.7507
```

4.3 Wilcox test

Následující Wilcoxonův testuje hypotézu, že střední hodnota nových případů na milión v ČR je rovna střední hodnotě nových případů na milión na Slovensku. Vzhledem ke skutečnosti, že p-value je menší než hladina významnosti ($\alpha = 0{,}05$), zamítáme tuto hypotézu ve prospěch alternativní. Při této hladině významnosti tudíž můžeme tvrdit, že střední hodnota nových případů na milión na Slovensku se nerovná střední hodnotě nových případů na milión v ČR. Vzhledem k zešikmení dat poskytuje tento test přesnější výsledky oproti dvouvýběrovému t-testu.

```
        Wilcoxon rank sum test with continuity correction

data:  new_cases_per_million_czechia and new_cases_per_million_slovakia
W = 205293, p-value = 2.97e-12
alternative hypothesis: true location shift is not equal to 0
```

Následující Wilcoxonův testuje hypotézu, že střední hodnota nových případů na milión v ČR je rovna střední hodnotě nových případů na milión v Německu. Vzhledem ke skutečnosti, že p-value je menší než hladina významnosti ($\alpha = 0{,}05$), zamítáme tuto hypotézu ve prospěch alternativní. Při této hladině významnosti tudíž můžeme tvrdit, že střední hodnota nových případů na milión v Německu se nerovná střední hodnotě nových případů na milión v ČR. Vzhledem k zešikmení dat poskytuje tento test přesnější výsledky oproti dvouvýběrovému t-testu.

```
        Wilcoxon rank sum test with continuity correction

data:  new_cases_per_million_czechia and new_cases_per_million_germany
W = 188720, p-value = 5.261e-05
alternative hypothesis: true location shift is not equal to 0
```

4.4 Fisherův test

Následující Fisherův test zkoumá zda jsou rozptyly hodnot nových případů na milión v ČR a na Slovensku stejné. Vzhledem ke skutečnosti, že p-value je menší než hladina významnosti ($\alpha = 0{,}05$), zamítáme tuto hypotézu ve prospěch alternativní, tudíž že jsou rozptyly těchto dat různé.

```
        F test to compare two variances


data:  new_cases_per_million_czechia and new_cases_per_million_slovakia
F = 4.5141, num df = 575, denom df = 575, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 3.832723 5.316656
sample estimates:
ratio of variances
          4.514119
```

4.5 Shapiro Wilk test

Následující Shapiro Wilk test testuje zda je veličina nových případů v ČR nabývá normálního rozdělení. Vzhledem ke skutečnosti, že p-value je menší než hladina významnosti ($\alpha = 0{,}05$), zamítáme tuto hypotézu ve prospěch alternativní, tudíž že tato veličina nenabývá normálního rozdělení.

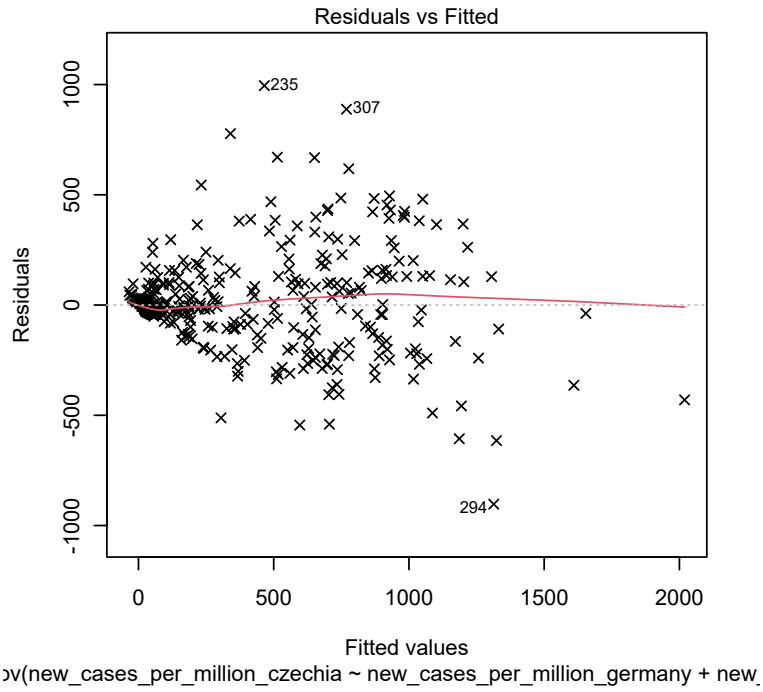
```
        Shapiro-Wilk normality test


data:  new_cases_czechia
W = 0.72003, p-value < 2.2e-16
```

Následující Shapiro Wilk test testuje zda je veličina nových testů v ČR nabývá normálního rozdělení. Vzhledem ke skutečnosti, že p-value je menší než hladina významnosti ($\alpha = 0{,}05$), zamítáme tuto hypotézu ve prospěch alternativní, tudíž že tato veličina nenabývá normálního rozdělení.

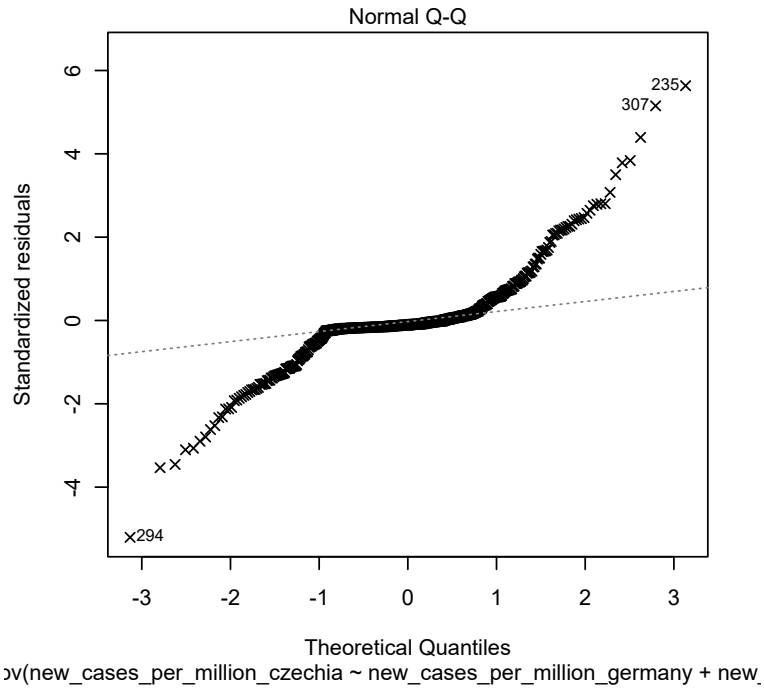
```
        Shapiro-Wilk normality test

data:  new_tests_czechia
W = 0.82915, p-value < 2.2e-16
```

5 ANOVA

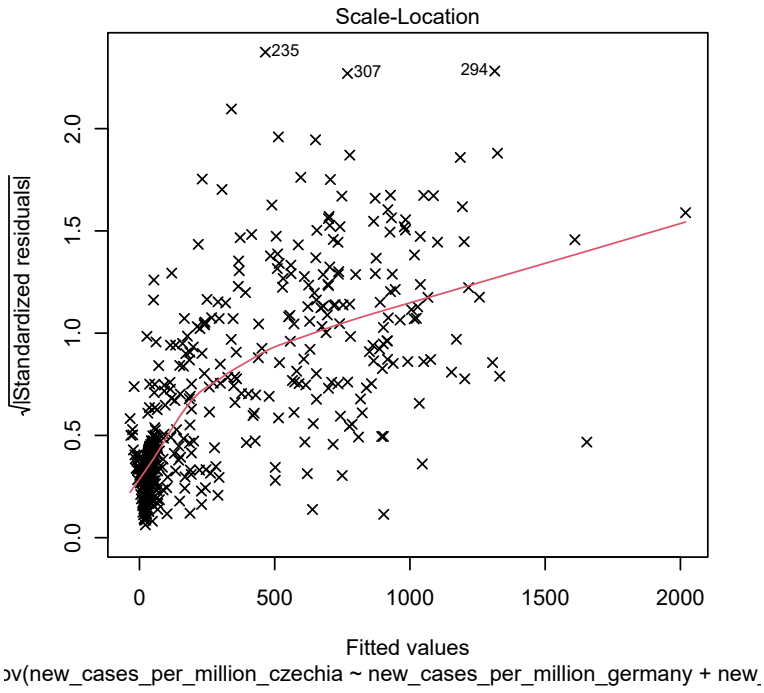
```
                                Df   Sum Sq  Mean Sq F value   Pr(>F)
new_cases_per_million_germany    1 32058325 32058325 1020.43  < 2e-16 ***
new_cases_per_million_slovakia   1 37842394 37842394 1204.54  < 2e-16 ***
new_cases_per_million_poland     1  2637470  2637470   83.95  < 2e-16 ***
new_cases_per_million_austria    1   978148   978148   31.14 3.72e-08 ***
Residuals                      571 17938795    31416
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



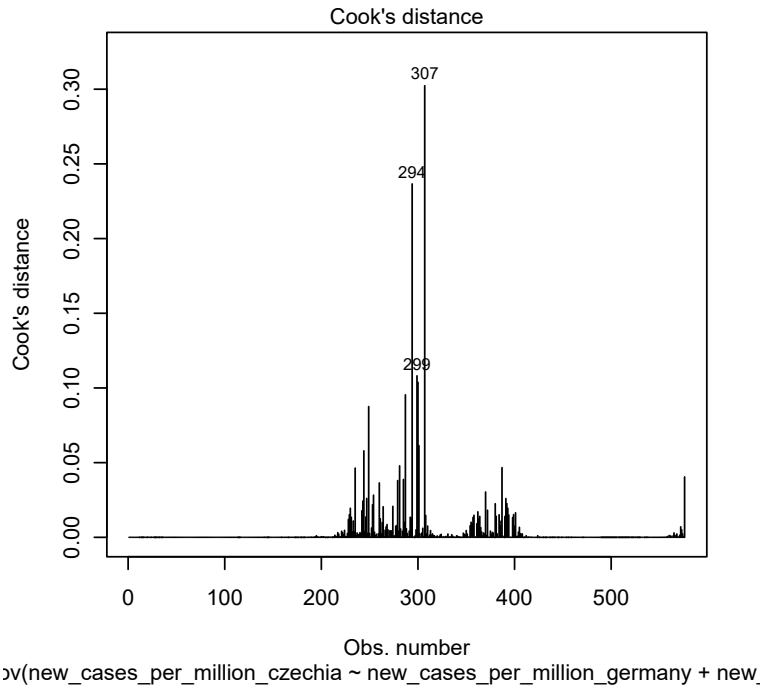
Obrázek 31: Anova graf nových testů, případů a úmrtí



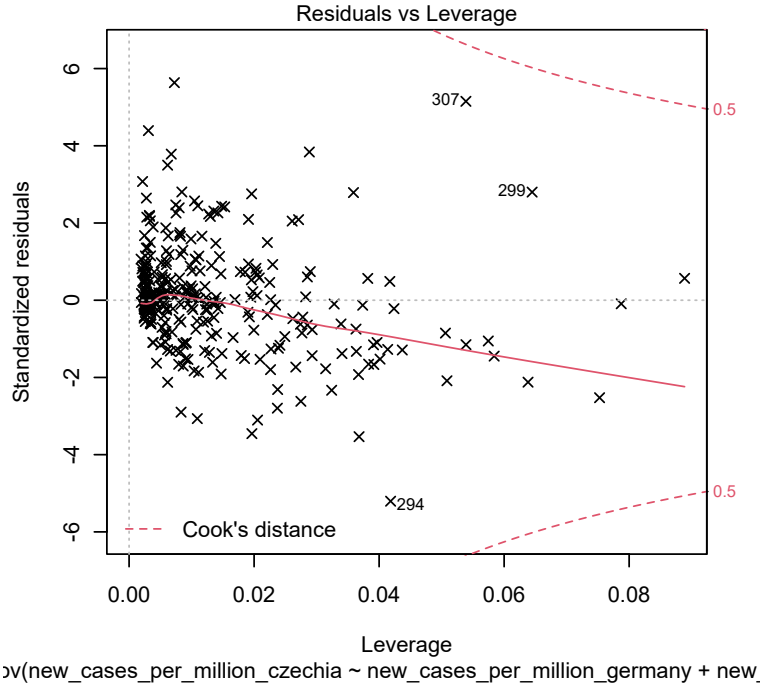
Obrázek 32: Anova graf nových testů, případů a úmrtí



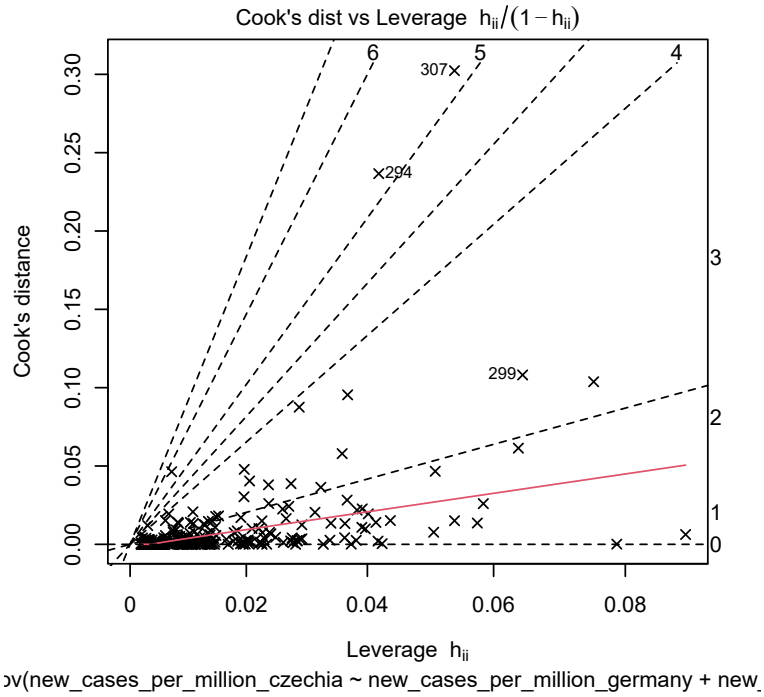
Obrázek 33: Anova graf nových testů, případů a úmrtí



Obrázek 34: Anova graf nových testů, případů a úmrtí



Obrázek 35: Anova graf nových testů, případů a úmrtí



Obrázek 36: Anova graf nových testů, případů a úmrtí

6 VARIANCE

Níže jsou popsány střední hodnoty kvadrátů odchylek od střední hodnoty nových testů a nových případů v ČR.

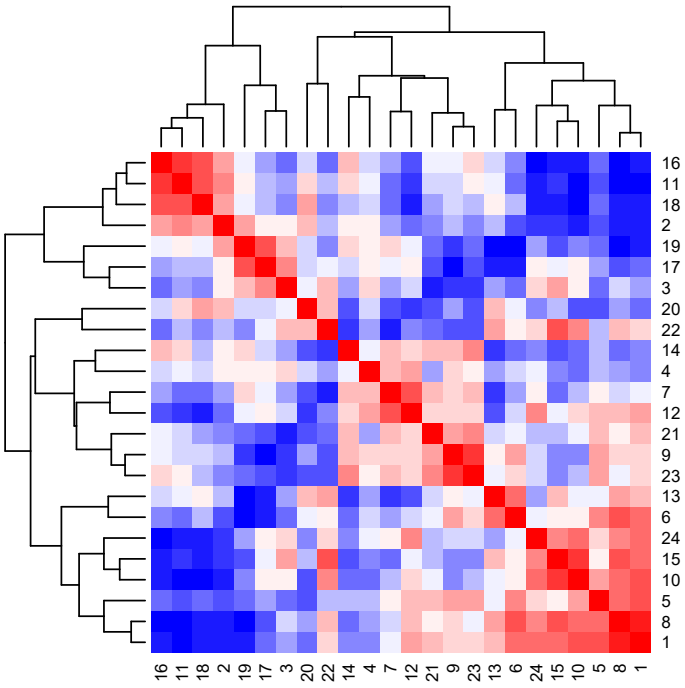
```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-61749419 -61749419 -61749419 -61749419 -61749419 -61749419
```

7 KORELACE

7.1 Korelační matice

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	1.00000000	-0.62274409	-0.129186615	-0.285154793	0.66579693	0.57478261
[2,]	-0.62274409	1.00000000	0.151838159	0.209449286	-0.43366681	-0.44705372
[3,]	-0.12918662	0.15183816	1.000000000	0.250435635	-0.36067001	-0.33492826
[4,]	-0.28515479	0.20944929	0.250435635	1.000000000	-0.10668415	-0.03613412
[5,]	0.66579693	-0.43366681	-0.360670011	-0.106684148	1.00000000	0.50445750
[6,]	0.57478261	-0.44705372	-0.334928261	-0.036134119	0.50445750	1.00000000
[7,]	0.05132667	-0.14357190	-0.163185379	0.386541958	0.15705896	-0.13049153
[8,]	0.89739130	-0.55838227	0.004784689	-0.171528226	0.57403785	0.69391304
[9,]	0.30782609	-0.09088933	-0.482383691	0.227688483	0.39791260	0.43652174
[10,]	0.72869565	-0.63448577	0.187472832	-0.359599784	0.45966516	0.17739130
[11,]	-0.66869565	0.54794522	-0.131361474	0.079233851	-0.46488368	-0.32956522
[12,]	0.42782609	-0.34007394	0.016093955	0.404005569	0.37442923	0.02869565
[13,]	0.38695652	-0.12002610	-0.172683792	-0.201567434	0.10784954	0.64347826
[14,]	-0.27478261	0.13785606	-0.133971305	0.135394108	-0.09306371	-0.34956522
[15,]	0.61043478	-0.47358122	0.410613349	-0.280365934	0.17612525	0.16434783
[16,]	-0.59304348	0.45749077	-0.321879108	0.026121050	-0.31919983	-0.23913043
[17,]	-0.31652174	0.21743858	0.533710359	0.180235243	-0.17090672	-0.62434783
[18,]	-0.59478261	0.47271147	-0.263157920	0.048323942	-0.32398348	-0.08000000
[19,]	-0.60608696	0.46879758	0.355371934	0.154549544	-0.33529029	-0.68608696
[20,]	-0.34398783	0.31165724	0.099412663	-0.006749405	-0.46889952	0.13089802
[21,]	0.32347826	-0.23265928	-0.586341943	-0.168916121	0.30876278	0.09043478
[22,]	0.23483367	-0.11048282	0.382423329	-0.190289684	-0.11439756	0.20656665
[23,]	0.25701240	-0.24880383	-0.511638037	0.139995729	0.47498913	0.25179387
[24,]	0.59143293	-0.52522836	0.291929526	0.050729401	0.30230535	0.09088933

	[,7]	[,8]	[,9]	[,10]	[,11]
[1,]	0.05132667	0.897391304	0.3078260870	0.72869565	-0.668695652
[2,]	-0.14357190	-0.558382270	-0.0908893259	-0.63448577	0.547945218
[3,]	-0.16318538	0.004784689	-0.4823836907	0.18747283	-0.131361474
[4,]	0.38654196	-0.171528226	0.2276884833	-0.35959978	0.079233851
[5,]	0.15705896	0.574037848	0.3979125991	0.45966516	-0.464883681
[6,]	-0.13049153	0.693913043	0.4365217391	0.17739130	-0.329565217
[7,]	1.00000000	-0.029578080	0.2801218186	-0.05480644	-0.334058318
[8,]	-0.02957808	1.000000000	0.3043478261	0.62956522	-0.664347826
[9,]	0.28012182	0.304347826	1.0000000000	-0.23304348	0.008695652
[10,]	-0.05480644	0.629565217	-0.2330434783	1.00000000	-0.646086957
[11,]	-0.33405832	-0.664347826	0.0086956522	-0.64608696	1.000000000
[12,]	0.68595048	0.310434783	0.2269565217	0.28782609	-0.530434783

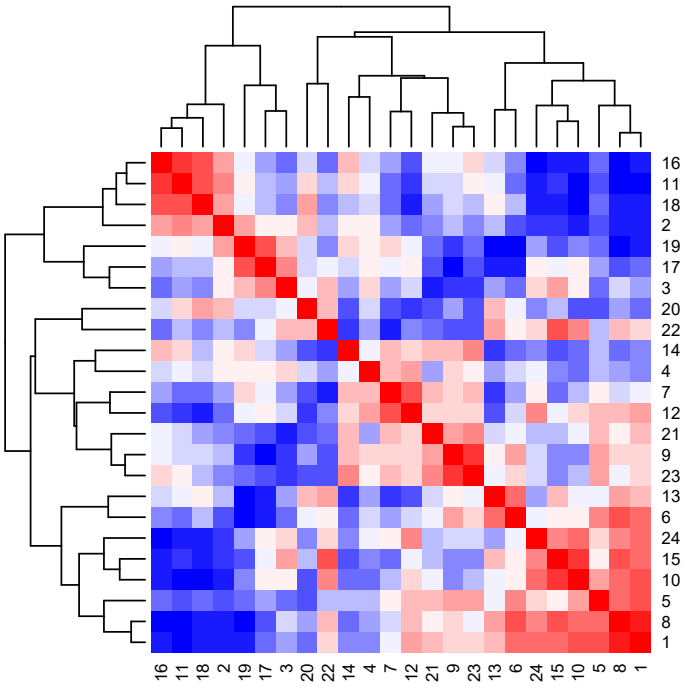


Obrázek 37: Heatmap graf korelační matice

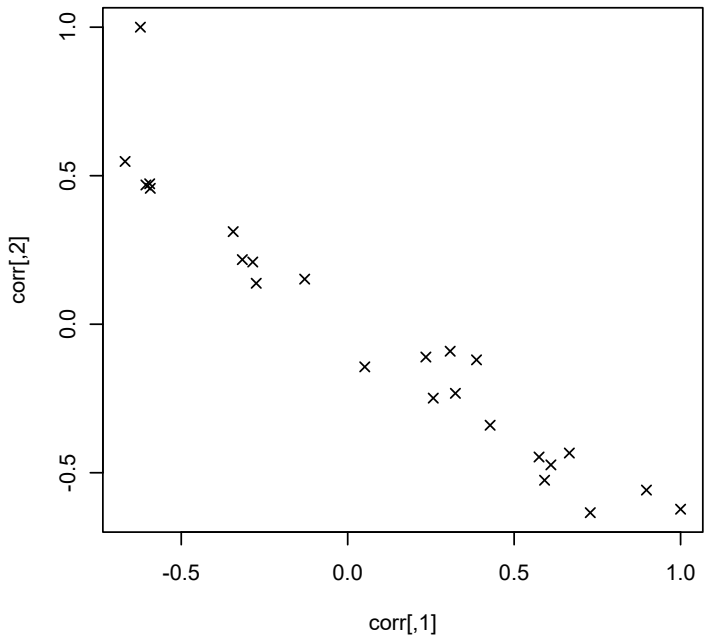
8 KOVARIANCE

8.1 Kovarianční matice

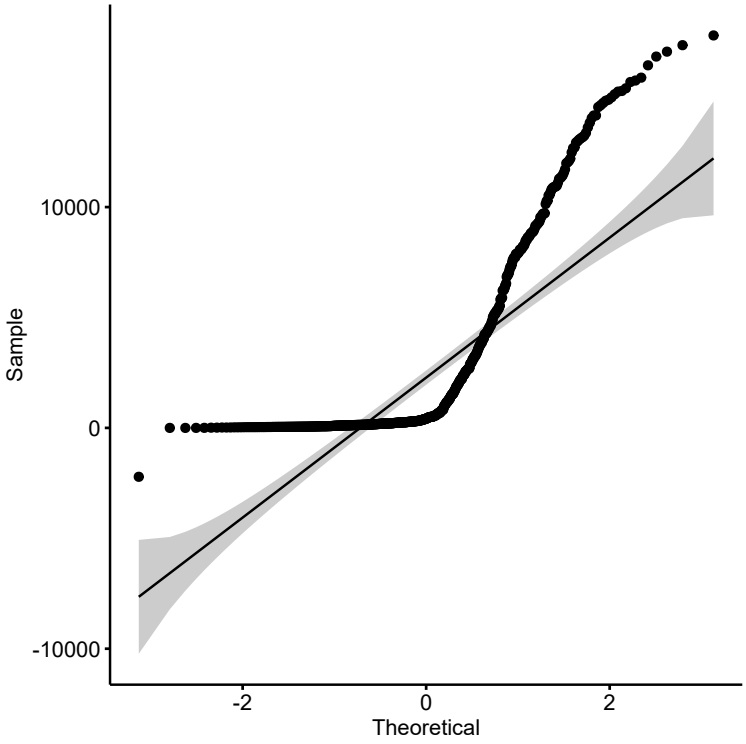
```
      V1                V2                V3                V4
Min.   :-33.435   Min.   :-31.717   Min.   :-29.3043   Min.   :-17.957
1st Qu.:-16.168   1st Qu.:-21.842   1st Qu.:-13.8859   1st Qu.: -8.467
Median : 12.293   Median : -5.761   Median : -3.1087   Median :  2.473
Mean   :  5.465   Mean   : -2.097   Mean   :  0.1916   Mean   :  3.088
3rd Qu.: 28.946   3rd Qu.: 12.046   3rd Qu.: 13.0217   3rd Qu.:  9.364
Max.   : 50.000   Max.   : 49.978   Max.   : 49.9565   Max.   : 49.870
      V5                V6                V7                V8
Min.   :-23.435   Min.   :-34.304   Min.   :-29.1739   Min.   :-33.217
1st Qu.:-16.016   1st Qu.:-13.087   1st Qu.:-10.8207   1st Qu.:-17.967
Median :  6.620   Median :  4.533   Median :  0.5435   Median :  8.315
Mean   :  4.815   Mean   :  3.619   Mean   :  2.1472   Mean   :  5.156
3rd Qu.: 20.663   3rd Qu.: 14.897   3rd Qu.: 14.7663   3rd Qu.: 25.891
Max.   : 49.978   Max.   : 50.000   Max.   : 49.9565   Max.   : 50.000
      V9               V10               V11               V12
Min.   :-34.435   Min.   :-36.435   Min.   :-33.435   Min.   :-31.174
1st Qu.:-10.658   1st Qu.:-17.337   1st Qu.:-23.321   1st Qu.:-14.668
Median :  8.826   Median :  6.196   Median : -1.522   Median : 10.457
Mean   :  4.632   Mean   :  3.081   Mean   : -1.861   Mean   :  4.583
3rd Qu.: 16.087   3rd Qu.: 23.522   3rd Qu.:  9.120   3rd Qu.: 16.321
Max.   : 50.000   Max.   : 50.000   Max.   : 50.000   Max.   : 50.000
     V13               V14               V15               V16
Min.   :-35.739   Min.   :-25.9348   Min.   :-30.8696   Min.   :-33.783
1st Qu.: -8.989   1st Qu.:-17.0652   1st Qu.:-17.2446   1st Qu.:-17.913
Median :  3.685   Median : -3.0435   Median :  0.4348   Median : -5.250
Mean   :  2.385   Mean   :  0.5525   Mean   :  2.7237   Mean   : -2.403
3rd Qu.: 17.924   3rd Qu.: 14.1957   3rd Qu.: 21.7337   3rd Qu.:  8.435
Max.   : 50.000   Max.   : 50.0000   Max.   : 50.0000   Max.   : 50.000
     V17               V18               V19               V20
Min.   :-34.4348   Min.   :-36.435   Min.   :-35.739   Min.   :-27.087
1st Qu.:-17.0543   1st Qu.:-19.696   1st Qu.:-19.120   1st Qu.:-19.397
Median : -0.6956   Median : -5.652   Median : -3.815   Median : -2.185
Mean   : -1.3895   Mean   : -3.584   Mean   : -3.570   Mean   : -2.415
3rd Qu.:  9.1413   3rd Qu.:  6.598   3rd Qu.:  8.679   3rd Qu.:  8.489
Max.   : 50.0000   Max.   : 50.000   Max.   : 50.000   Max.   : 49.978
     V21               V22               V23               V24
Min.   :-29.304   Min.   :-29.174   Min.   :-25.565   Min.   :-33.783
1st Qu.: -9.234   1st Qu.:-14.495   1st Qu.:-12.223   1st Qu.: -9.245
```



Obrázek 38: Heatmap graf kovarianční matice



Obrázek 39: Graf kovarianční matice



Obrázek 40: GGQQPlot graf korelační matice

9 TESTOVÁNÍ V KONTINGENČNÍCH TABULKÁCH

9.1 Pearsonův Chí-kvadrát test

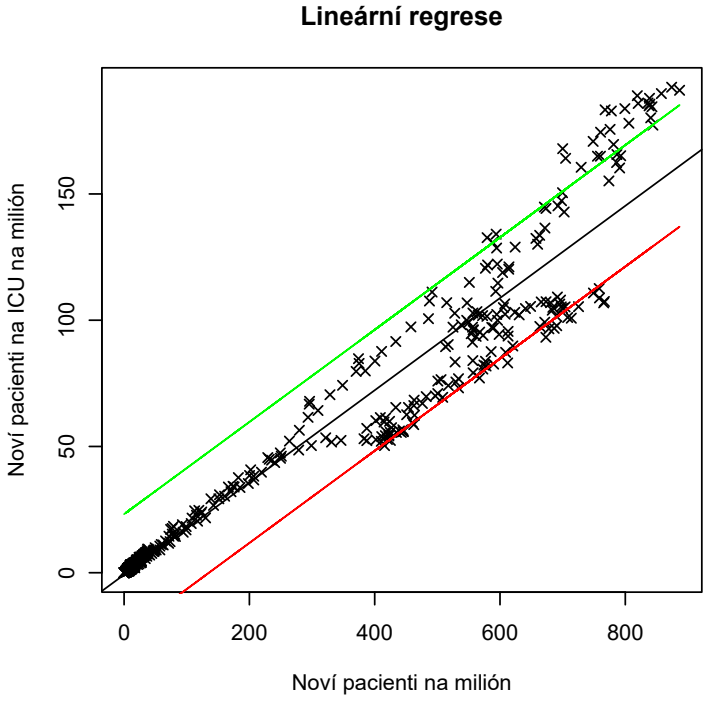
Následující chí-kvadrát test zkoumá zda má veličina nových případů v ČR stejné rozdělení jako veličina nových případů v ČR. Vzhledem ke skutečnosti, že p-value je vyšší než hladina významnosti ($\alpha = 0{,}05$), tuto hypotézu nemůžeme zamítnout.

```
        Pearson's Chi-squared test

data:  new_tests_czechia and new_cases_czechia
X-squared = 147356, df = 146982, p-value = 0.245
```

10 REGRESE

10.1 Lineární regrese

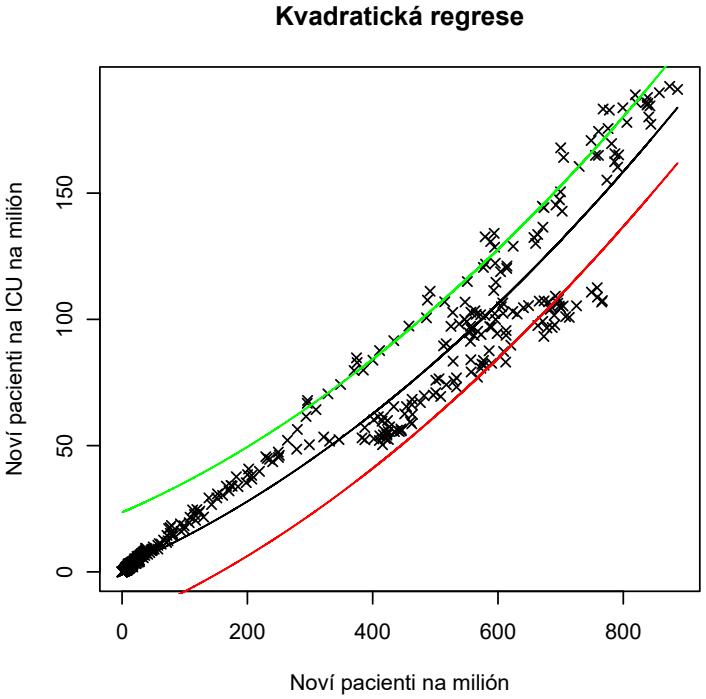
Následující graf zobrazuje jakých hodnot bude s 95% pravděpodobností nabývat hodnota pacientů na ICU na milión v ČR v závislosti na počtu nově hospitalizovaných pacientů na milión v ČR. Tato závislost je zde vyjádřena jako lineární funkce $y = -0.7746 + 0.1826x$.



Obrázek 41: Graf lineární regrese

10.2 Kvadratická regrese

Následující graf zobrazuje jakých hodnot bude s 95% pravděpodobností nabývat hodnota pacientů na ICU na milión v ČR v závislosti na počtu nově hospitalizovaných pacientů na milión v ČR. Tato závislost je zde vyjádřena jako kvadratická funkce $y = 1.9982244 + 0.1074610x + 0.0001102x^2$.



Obrázek 42: Graf kvadraditcké regrese

ZÁVĚR

V této semestrální práci byl analyzován vývoj epidemie nemoci Covid-19 v ČR. Jak je zřejmé z grafů, zkoumaná data se neřídí dle normálního rozdělení pravděpodobnosti a zpravidla jsou výrazně zešikmena vlevo. Pomocí grafů byly porovnány sedmidenní klouzavé průměry nových případů na milión v Rakousku a České republice. Pomocí grafů byly také vizualizovány další veličiny jako například počet nových hospitalizací, počet nových testů, reprodukční číslo, počet pacientů na ICU či pozitivita testů. Pomocí testů bylo otestováno například zda se střední hodnota nových případů rovná konkrétní hodnotě či zda se střední hodnota nových případů výrazněji v průběhu času změnila. Tyto analýzy byly proté vzhledem k sešikmení dat provedeny kromě t-testu také pomocí Wilcox testu, jelikož by jeho výsledky zešikmení dat nemělo případně tolik ovlivnit. Pro řádné srovnání nových případů na milión je poté na data aplikován test ANOVA, který tuto veličinu porovnává mezi ČR, Německem, Slovenskem, Polskem a Rakouskem. Nakonec je pomocí regrese navržena lineární a kvadratická funkce popisující možnou závislost přírůstku nových pacientů na ICU na milión na přírustku nově hospitalizovaných pacientů na milión. Těmito operacemi práce jistě poskytuje bližší pohled na vývoj současné epidemie v naší zemi jakož i nebezpečí, které tento virus představuje.

POUŽITÁ LITERATURA

[1] Our World in Data *Data on COVID-19 (coronavirus)* [online]. 2021 [cit. 2021-11-18]. Dostupné z: `https://github.com/owid/covid-19-data/tree/master/public/data`

SEZNAM PŘÍLOH

PŘÍLOHA A

Příloha A zahrnuje ZIP soubor, který obsahuje:

- Zdrojové kódy
- Zdrojová data použitá v práci