Zápočtová úloha NMAI061/2019

Jméno a příjmení

Datum narození: 19700131

1. V hlavičce tohoto souboru správně uveďte svoje jméno (v příkazu `\author{...}`) a datum narození (`\date{...}`) ve formátu `yyyymmdd`.
2. TEXový kód zadání Vaší zápočtové úlohy získáte pomocí příkazu `Sweave("zapoctovka2019.Rnw")` v R. Přitom je nutné mít nainstalovanou knihovnu `lmtest`. Příkaz `Sweave("zapoctovka2019.Rnw")` zadávejte až po zadání svého data narození do souboru `zapoctovka2019.Rnw`!
3. Vzniklý TEXový soubor `zapoctovka.tex` přejmenujte na `prijmenijmeno.tex` a přeložte do PDF pomocí pdflatexu.
4. Odpovědi na otázky můžete psát přímo do TEXového souboru nebo do zvláštního souboru `prijmenijmeno-odpovedi.pdf`.
5. Úlohu můžete odevzdat elektronicky (buď soubor prijmenijmeno.pdf se zadáním i s odpověďmi nebo soubor prijmenijmeno.pdf se zadáním a soubor prijmenijmeno-odpovedi.pdf s odpověďmi).
6. Za každou správně vyřešenou úlohu získáte 1 bod. Pro získání zápočtu je nutné získat minimálně 8 bodů (počet bodů může být zohledněn i při zkoušce).

Data

Načtěte datový soubor `Loblolly` a nastudujte jeho popis v nápovědě.

```
> data(Loblolly)
```

Analyzovat budeme data pouze pro dva náhodně vybrané stromy. Výsledná data jsou:

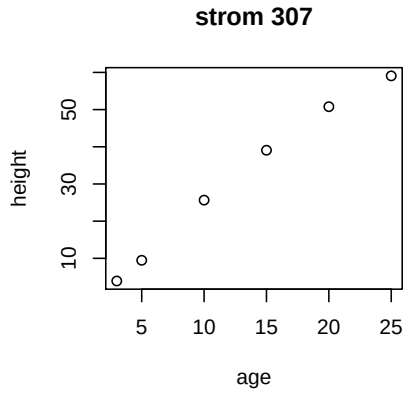
```
   height age Seed
1    3.91   3  307
2    9.48   5  307
3   25.66  10  307
4   39.07  15  307
5   50.78  20  307
6   59.07  25  307
7    3.93   3  329
8    9.34   5  329
9   26.08  10  329
10  37.79  15  329
11  48.31  20  329
12  56.43  25  329
```

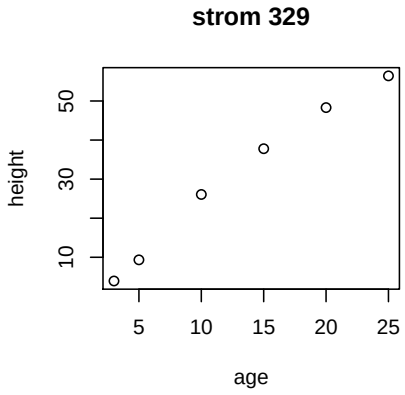
Vaším úkolem bude prozkoumat závislost výšky těchto dvou stromů na jejich věku.

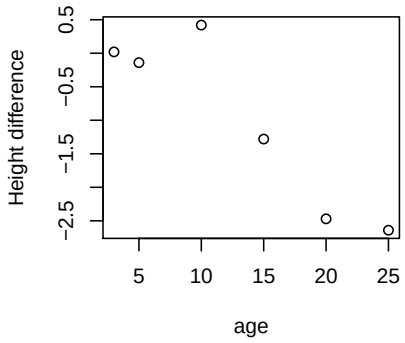
Úkoly

Úkol 1: Okomentujte následující popisné statistiky a grafy. Mimo jiné můžete okomentovat, který strom je vyšší, vysvětlit význam jednotlivých popisných statistik (tj. co je z těchto hodnot vidět) a vysvětlit proč jsou statistiky u proměnné age stejné pro oba stromy.

```
> sapply(stromy[,-3],tapply,Seed,mean)

      height age
307 31.32833  13
329 30.31333  13

> sapply(stromy[,-3],tapply,Seed,median)

    height  age
307 32.365 12.5
329 31.935 12.5

> sapply(stromy[,-3],tapply,Seed,sd)

      height      age
307 22.22227 8.602325
329 21.05109 8.602325

> sapply(stromy[,-3],tapply,Seed,var)

      height age
307 493.8291  74
329 443.1486  74

> par(mfrow=c(2,2))
> plot(height~age,main=paste("strom",Seed[k1s]),subset=strom1)
> plot(height~age,main=paste("strom",Seed[k2s]),subset=strom2)
> hist(age)
> poc=max(k)-1
> plot(height[strom2]-height[strom1]~age[strom1],ylab="Height difference",xlab="age",pch=1,col=
>
```







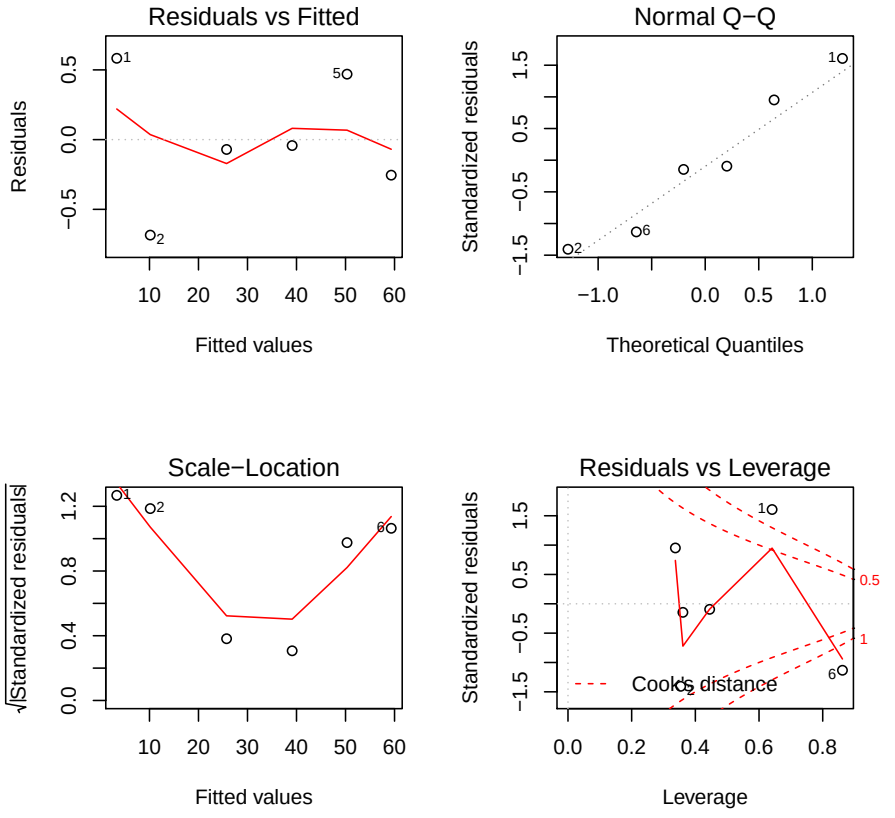
Úkol 1: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v TEXu) nebo můžete každé řešení označit číslem úkolu a vše odevzdat ve zvláštním dokumentu.

Úkol 2: Vysvětlete, co přesně znamenají jednotlivé odhady koeficientů v následujícím lineárním modelu. Které koeficienty jsou významně odlišné od nuly?

```
> strom1.lm=lm(height~age+I(age^2),subset=strom1)
> strom1.lm

Call:
lm(formula = height ~ age + I(age^2), subset = strom1)

Coefficients:
(Intercept)          age     I(age^2)
   -7.58511      3.76834     -0.04368

> summary(strom1.lm)

Call:
lm(formula = height ~ age + I(age^2), subset = strom1)

Residuals:
       1        2        3        4        5        6
 0.58318 -0.68467 -0.07054 -0.04253  0.46938 -0.25482
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.585107   0.809513  -9.370 0.002575 **
age          3.768344   0.145436  25.911 0.000126 ***
I(age^2)    -0.043678   0.005171  -8.447 0.003482 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6058 on 3 degrees of freedom
Multiple R-squared:  0.9996,        Adjusted R-squared:  0.9993
F-statistic:  3363 on 2 and 3 DF,  p-value: 9.415e-06
```

```
> par(mfrow=c(2,2))
> plot(strom1.lm)
```

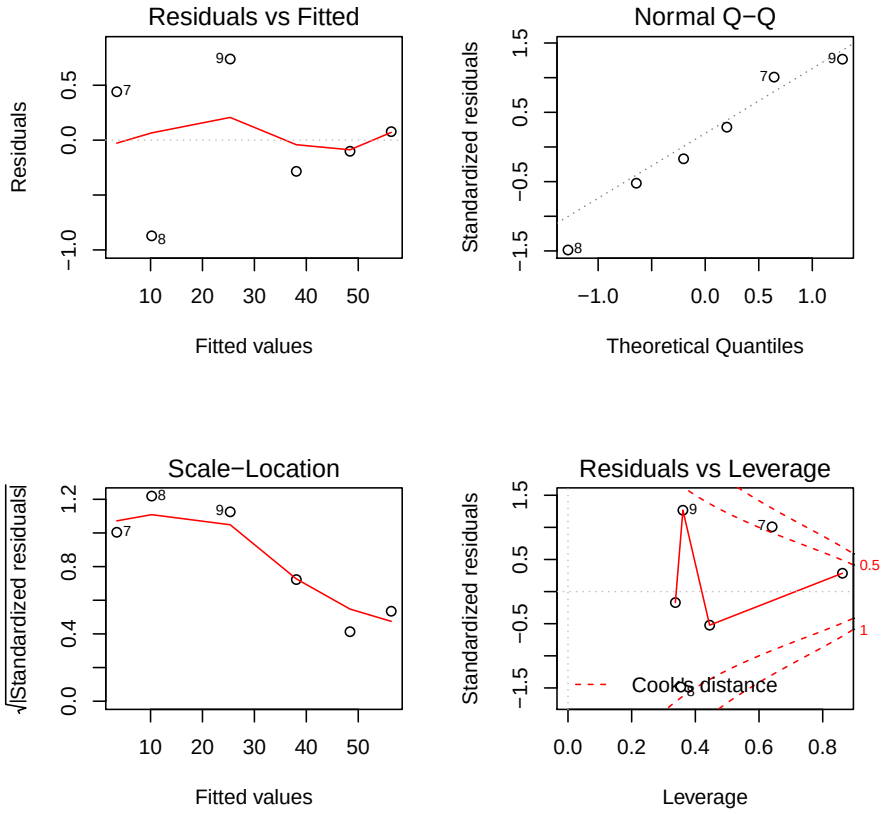


Úkol 2: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v TEXu) nebo můžete každé řešení označit číslem úkolu a vše odevzdat ve zvláštním dokumentu.

Úkol 3: Vysvětlete, co přesně znamenají jednotlivé odhady koeficientů v následujícím lineárním modelu. Které koeficienty jsou významně odlišné od nuly?

```
> strom2.lm=lm(height~age+I(age^2),subset=strom2)
> strom2.lm
```

```
Call:
lm(formula = height ~ age + I(age^2), subset = strom2)

Coefficients:
(Intercept)          age     I(age^2)
   -7.31358      3.74467     -0.04792

> summary(strom2.lm)

Call:
lm(formula = height ~ age + I(age^2), subset = strom2)

Residuals:
       7        8        9       10       11       12
 0.44086 -0.87174  0.73899 -0.28422 -0.10139  0.07749

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.313581   0.975139  -7.500 0.004911 **
age          3.744669   0.175193  21.375 0.000224 ***
I(age^2)    -0.047921   0.006229  -7.694 0.004563 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7297 on 3 degrees of freedom
Multiple R-squared:  0.9993,        Adjusted R-squared:  0.9988
F-statistic:  2079 on 2 and 3 DF,  p-value: 1.936e-05

> par(mfrow=c(2,2))
> plot(strom2.lm)
```



Úkol 3: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v TEXu) nebo můžete každé řešení označit číslem úkolu a vše odevzdat ve zvláštním dokumentu.

Úkol 4: Pomocí následujícího lineárního modelu rozhodněte, jestli oba stromy mohly být při zasazení stejně vysoké a jestli oba stromy rostly stejně rychle. Který strom roste rychleji? Je pozorovaný rozdíl rychlosti růstu statisticky signifikantní? Při interpretaci parametrů zvažte i matici modelu získanou příkazem `model.matrix(strom.lm)`.

```
> strom.lm=lm(height~age+I(age^2)+Seed+age:Seed)
> summary(strom.lm)

Call:
lm(formula = height ~ age + I(age^2) + Seed + age:Seed)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90146 -0.17155 -0.08866  0.35164  0.84144

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.853074   0.696856 -11.269 9.68e-06 ***
age          3.826601   0.110286  34.697 4.28e-09 ***
I(age^2)    -0.045799   0.003832 -11.951 6.53e-06 ***
Seed329      0.807459   0.708973   1.139   0.2922
```

```
age:Seed329 -0.140189   0.046681  -3.003   0.0199 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6349 on 7 degrees of freedom
Multiple R-squared:  0.9994,        Adjusted R-squared:  0.9991
F-statistic:  2905 on 4 and 7 DF,  p-value: 2.407e-11
```

```
> model.matrix(strom.lm)

   (Intercept) age I(age^2) Seed329 age:Seed329
1            1   3        9       0           0
2            1   5       25       0           0
3            1  10      100       0           0
4            1  15      225       0           0
5            1  20      400       0           0
6            1  25      625       0           0
7            1   3        9       1           3
8            1   5       25       1           5
9            1  10      100       1          10
10           1  15      225       1          15
11           1  20      400       1          20
12           1  25      625       1          25
attr(,"assign")
[1] 0 1 2 3 4
attr(,"contrasts")
attr(,"contrasts")$Seed
[1] "contr.treatment"
```

```
> anova(strom.lm2<-lm(height~age+I(age^2)),strom.lm)

Analysis of Variance Table

Model 1: height ~ age + I(age^2)
Model 2: height ~ age + I(age^2) + Seed + age:Seed
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1      9 9.5484
2      7 2.8219  2    6.7265 8.3428 0.01403 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Úkol 4: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v TEXu) nebo můžete každé řešení označit číslem úkolu a vše odevzdat ve zvláštním dokumentu.

Úkol 5: Je některé z následujících studentizovaných reziduí příliš velké?

```
> rstudent(strom.lm)

          1           2           3           4           5           6
 1.88438818 -1.34896573 -0.29863607 -0.30283860  0.77990901 -0.31970672
          7           8           9          10          11          12
 0.73687377 -2.17499856  1.78369945 -0.27525311 -0.09338353 -0.16259143
```

Úkol 5: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v TEXu) nebo můžete každé řešení označit číslem úkolu a vše odevzdat ve zvláštním dokumentu.

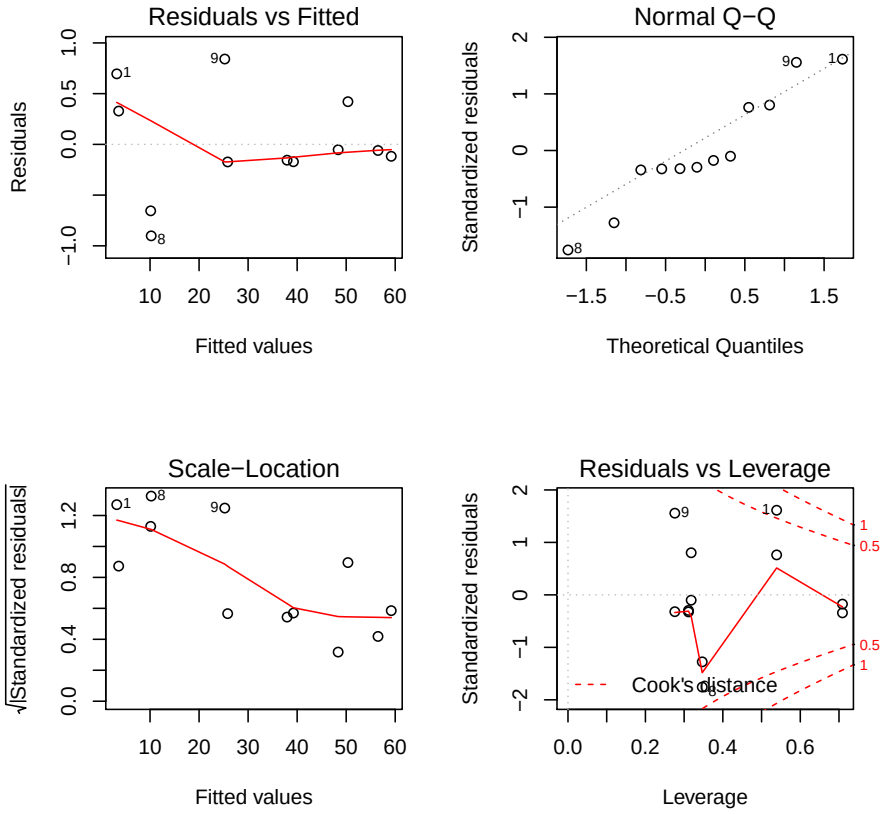
Úkol 6: Rozhodněte, jestli je některá z následujících hodnot „leverage“ „podezřelá“.

```
> hatvalues(strom.lm)

        1         2         3         4         5         6         7         8
0.5389887 0.3467890 0.2759353 0.3111953 0.3181744 0.7089174 0.5389887 0.3467890
        9        10        11        12
0.2759353 0.3111953 0.3181744 0.7089174

> influence.measures(strom.lm)

Influence measures of
         lm(formula = height ~ age + I(age^2) + Seed + age:Seed) :

    dfb.1_ dfb.age dfb.I..2 dfb.S329 dfb.a.S3   dffit cov.r   cook.d   hat inf
1   1.9257 -1.2778   0.8866 -1.28752   1.0202  2.0375 0.459 0.608544 0.539   *
2  -0.7789  0.3424  -0.1411  0.66927  -0.4908 -0.9829 0.880 0.172962 0.347
3  -0.0160 -0.0812   0.1023  0.08551  -0.0387 -0.1844 2.773 0.007814 0.276
4   0.0606 -0.1387   0.1334  0.03150   0.0268 -0.2036 2.909 0.009522 0.311
5  -0.1588  0.2273  -0.1304  0.06706  -0.2430  0.5328 1.956 0.060134 0.318
6  -0.0234  0.1106  -0.2318 -0.13530   0.2614 -0.4989 6.824 0.057112 0.709   *
7   0.2408 -0.3308   0.3467  0.50347  -0.3990  0.7968 3.040 0.135833 0.539
8  -0.1581  0.2171  -0.2275 -1.07909   0.7914 -1.5848 0.181 0.327667 0.347   *
9  -0.4244  0.5829  -0.6109  0.51075  -0.2312  1.1011 0.356 0.184877 0.276
10  0.0842 -0.1157   0.1213 -0.02863  -0.0244 -0.1850 2.947 0.007887 0.311
11  0.0108 -0.0149   0.0156  0.00803  -0.0291 -0.0638 3.147 0.000948 0.318   *
12 -0.0819  0.1125  -0.1179  0.06881  -0.1329 -0.2537 7.264 0.014957 0.709   *
```

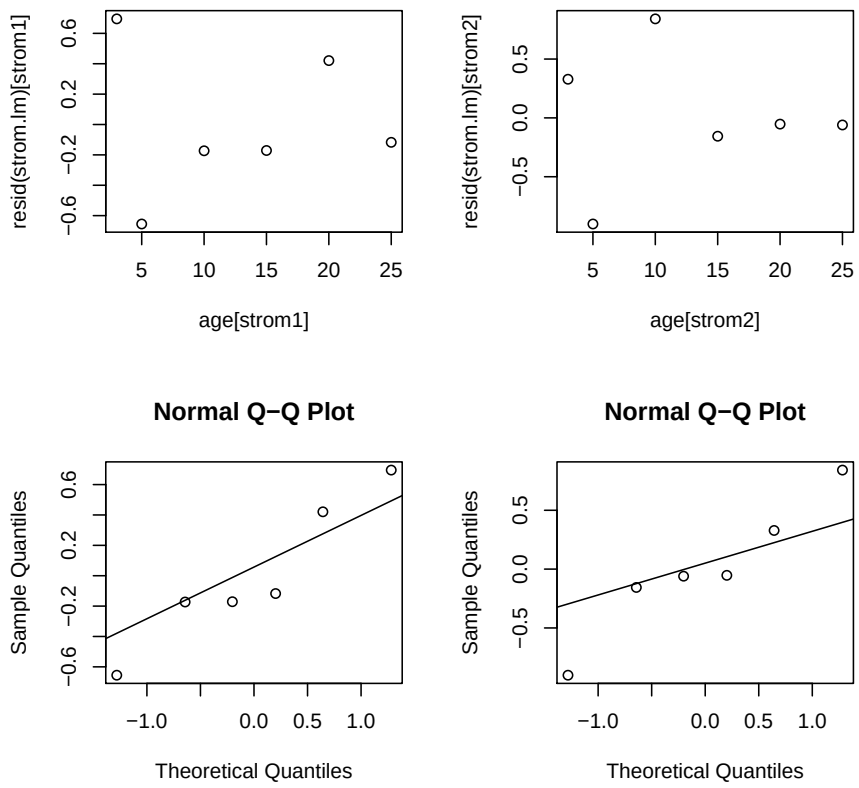
Úkol 6: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v TEXu) nebo můžete každé řešení označit číslem úkolu a vše odevzdat ve zvláštním dokumentu.

Úkol 7: Jsou podle Vašeho názoru splněny všechny předpoklady normálního lineárního modelu? Jak byste na následujících grafech poznali např. porušení předpokladu nezávislosti? Jaké důsledky by mělo porušení tohoto předpokladu?

```
>   par(mfrow=c(2,2))
>   plot(strom.lm)
```




```
> par(mfrow=c(2,2))
> plot(resid(strom.lm)[strom1]~age[strom1])
> plot(resid(strom.lm)[strom2]~age[strom2])
> qqnorm(resid(strom.lm)[strom1])
> qqline(resid(strom.lm)[strom1])
> qqnorm(resid(strom.lm)[strom2])
> qqline(resid(strom.lm)[strom2])
```



Úkol 7: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v TEXu) nebo můžete každé řešení označit číslem úkolu a vše odevzdat ve zvláštním dokumentu.

Úkol 8: Shapiro-Wilkův test se používá k ověřování normality. Okomentujte výsledek tohoto testu použitého na rezidua modelu.

```
> shapiro.test(resid(strom.lm))

        Shapiro-Wilk normality test

data:  resid(strom.lm)
W = 0.94514, p-value = 0.5674

> shapiro.test(rstandard(strom.lm))

        Shapiro-Wilk normality test

data:  rstandard(strom.lm)
W = 0.92629, p-value = 0.3424

> shapiro.test(rstudent(strom.lm))

        Shapiro-Wilk normality test

data:  rstudent(strom.lm)
W = 0.92825, p-value = 0.362
```

Úkol 8: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v TEXu) nebo můžete každé řešení označit číslem úkolu a vše odevzdat ve zvláštním dokumentu.

Úkol 9: Durbin-Watsonův test se používá k ověřování nezávislosti po sobě jdoucích pozorování. Okomentujte výsledek tohoto testu použitého na rezidua modelu.

```
> library(lmtest)
> dwtest(strom.lm)

        Durbin-Watson test

data:  strom.lm
DW = 2.9945, p-value = 0.6872
alternative hypothesis: true autocorrelation is greater than 0
```

Úkol 9: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v TEXu) nebo můžete každé řešení označit číslem úkolu a vše odevzdat ve zvláštním dokumentu.

Úkol 10: Popište, jak byste otestovali hypotézu, že oba stromy byly ve věku 14 let stejně vysoké?

Úkol 10: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v TEXu) nebo můžete každé řešení označit číslem úkolu a vše odevzdat ve zvláštním dokumentu.

Úkol 11: Stručně a výstižně shrňte výsledky získané analýzou lineárního modelu `strom.lm`, a vysvětlete, které předpoklady normálního lineárního modelu jsou v tomto případě porušené (pokud tam nějaké takové jsou) a navrhněte vhodný způsob, jakým by bylo možné model `strom.lm` dále vylepšit.

Úkol 11: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v TEXu) nebo můžete každé řešení označit číslem úkolu a vše odevzdat ve zvláštním dokumentu.

Jenom pro kontrolu: datum narození je 19700131 a vybrali jsme stromy 4 a 1.