

Zápočtová úloha NMAI061/2017

Jméno a příjmení

Datum narození: 19700131

1. V hlavičce tohoto souboru správně uveďte svoje jméno (v příkazu `\author{...}`) a datum narození (`\date{...}`).
2. T_EXový kód zadání Vaší zápočtové úlohy získáte pomocí příkazu `Sweave("zapoctovka2017.Rnw")` v R. Přitom je nutné mít nainstalovanou knihovnu `lmtest`. Příkaz `Sweave("zapoctovka2017.Rnw")` zadávejte až po zadání svého data narození do souboru `zapoctovka2017.Rnw!`
3. Vzniklý T_EXový soubor `zapoctovka.tex` přejmenujte na `prijmeni.jmeno.tex` a přeložte do PDF pomocí `pdflatex` (nebo do `dvi` pomocí `latex` a pak do PDF pomocí `dvipdf`).
4. Odpovědi na otázky můžete psát přímo do Vašeho T_EXového souboru nebo do zvláštního souboru `prijmeni.jmeno-odpovedi.pdf`.
5. Úlohu můžete odevzdat elektronicky (buď soubor `prijmeni.jmeno.pdf` se zadáním i s odpověďmi nebo soubor `prijmeni.jmeno.pdf` se zadáním a soubor `prijmeni.jmeno-odpovedi.pdf` s odpověďmi).
6. Za každou správně vyřešenou úlohu získáte 1 bod. Pro získání zápočtu je nutné získat minimálně 7 bodů (počet bodů může být zohledněn i při zkoušce).
7. Zápočet je po dohodě s příslušným cvičícím možné zapsat i při zkoušce.

Data

Načtete datový soubor `pig.weights.txt` a nastudujte jeho popis v nápovědě.

```
> PigWeight=read.table("http://www.karlin.mff.cuni.cz/~hlavka/vyuka/stati/zapoctovka/data/pig.weight.txt")
> names(PigWeight)=c("ID", "Time", "weight")
```

Analyzovat budeme data pouze pro dvě náhodně vybraná prasata. Výsledná data jsou:

	ID	Time	weight
	100	12	1 24.0
	101	12	2 29.0
	102	12	3 39.0
	103	12	4 44.0
	104	12	5 50.5
	105	12	6 57.0
	106	12	7 61.5
	107	12	8 68.0
	108	12	9 73.5
	154	18	1 27.5
	155	18	2 32.5
	156	18	3 36.0
	157	18	4 43.0
	158	18	5 49.5

159	18	6	52.5
160	18	7	56.0
161	18	8	61.0
162	18	9	64.0

Vaším úkolem bude prozkoumat závislost váhy těchto dvou prasat na čase.

Úkoly

Úkol 1: Okomentujte následující popisné statistiky a grafy (mimo jiné můžete okomentovat, které prase je těžší, vysvětlit význam jednotlivých popisných statistik, tj. co tyto statistiky odhadují nebo měří a jaké jsou jejich vztahy a vysvětlit proč jsou statistiky u proměnné Time stejné pro obě prasata).

```
> sapply(prasata[, -1], tapply, ID, mean)

      Time  weight
12      5 49.61111
18      5 46.88889

> sapply(prasata[, -1], tapply, ID, median)

      Time weight
12      5  50.5
18      5  49.5

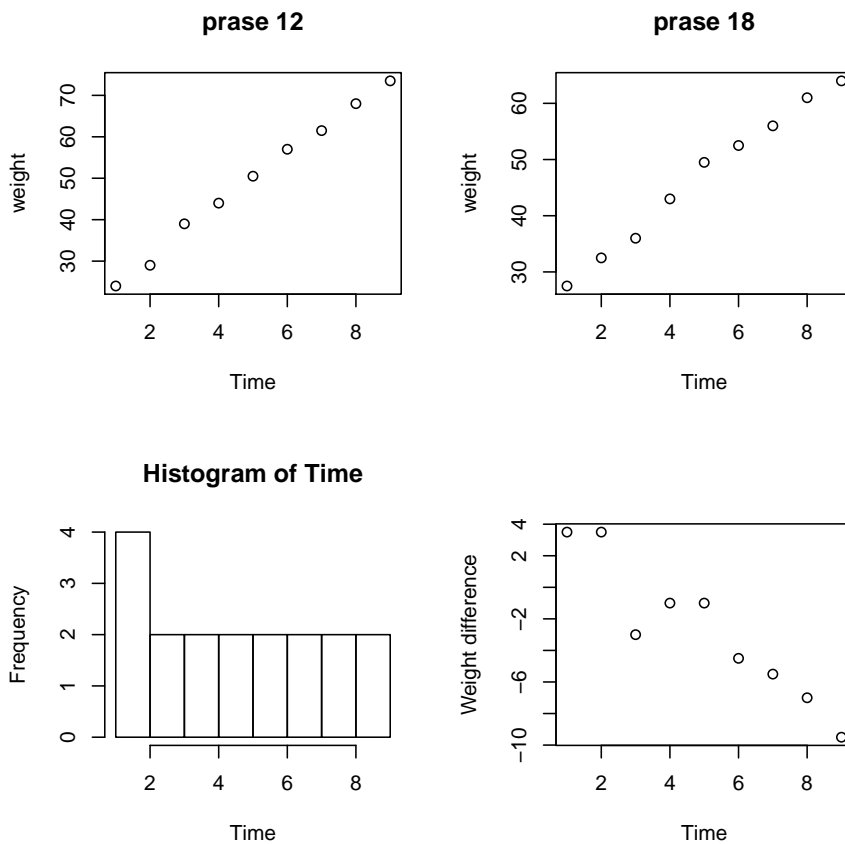
> sapply(prasata[, -1], tapply, ID, sd)

      Time  weight
12 2.738613 17.06747
18 2.738613 12.89326

> sapply(prasata[, -1], tapply, ID, var)

      Time  weight
12  7.5 291.2986
18  7.5 166.2361

> par(mfrow=c(2,2))
> plot(weight~Time, main=paste("prase", ID[1]), subset=(ID==ID[1]))
> plot(weight~Time, main=paste("prase", ID[13]), subset=(ID==ID[13]))
> hist(Time)
> poc=max(k)-1
> plot(weight[10:18]-weight[1:9]~Time[1:9], ylab="Weight difference", xlab="Time")
```



Úkol 1: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v $\text{T}_{\text{E}}\text{X}$ u) 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?

```
> prase1.lm=lm(weight~Time+I(Time^2),subset=(ID==ID[k1s]))
> prase1.lm
```

Call:

```
lm(formula = weight ~ Time + I(Time^2), subset = (ID == ID[k1s]))
```

Coefficients:

(Intercept)	Time	I(Time^2)
16.107	7.537	-0.132

```
> summary(prase1.lm)
```

Call:

```
lm(formula = weight ~ Time + I(Time^2), subset = (ID == ID[k1s]))
```

Residuals:

Min	1Q	Median	3Q	Max
-1.65303	-0.14264	0.04697	0.42403	1.47013

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	16.10714	1.29374	12.450	1.64e-05	***
Time	7.53701	0.59404	12.688	1.47e-05	***
I(Time^2)	-0.13203	0.05794	-2.279	0.0629	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

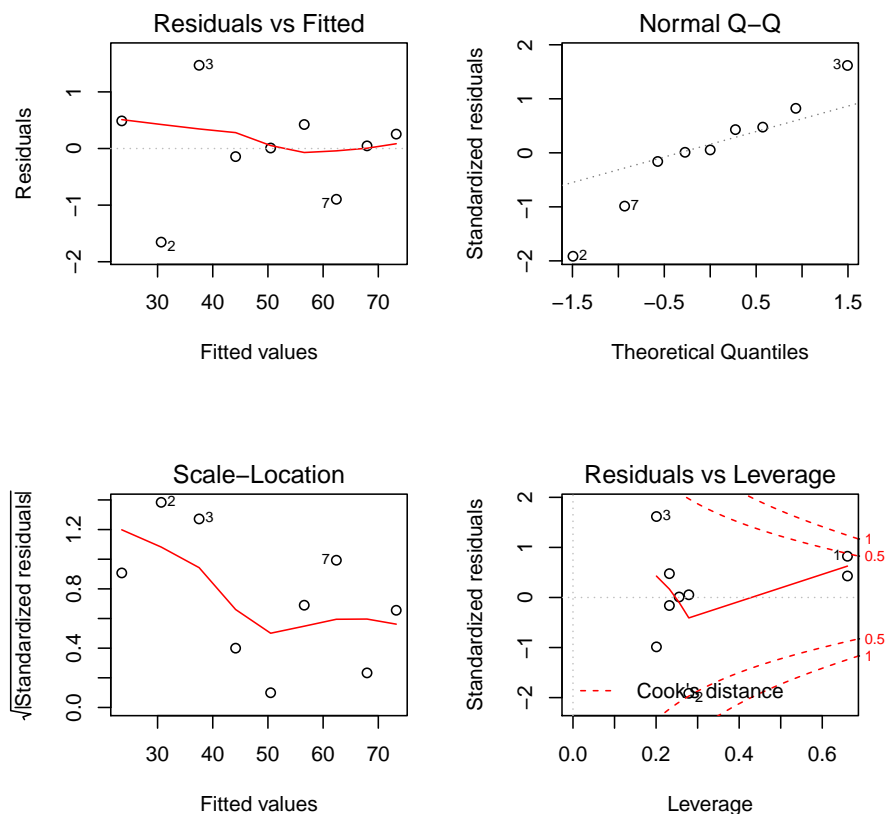
Residual standard error: 1.017 on 6 degrees of freedom

Multiple R-squared: 0.9973, Adjusted R-squared: 0.9965

F-statistic: 1124 on 2 and 6 DF, p-value: 1.886e-08

```
> par(mfrow=c(2,2))
```

```
> plot(prase1.lm)
```



Úkol 2: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v \TeX u) 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?

```
> prase2.lm=lm(weight~Time+I(Time^2),subset=(ID==ID[k2s]))  
> prase2.lm
```

```

Call:
lm(formula = weight ~ Time + I(Time^2), subset = (ID == ID[k2s]))

Coefficients:
(Intercept)      Time      I(Time^2)
  20.8333      6.1227     -0.1439

> summary(prase2.lm)

Call:
lm(formula = weight ~ Time + I(Time^2), subset = (ID == ID[k2s]))

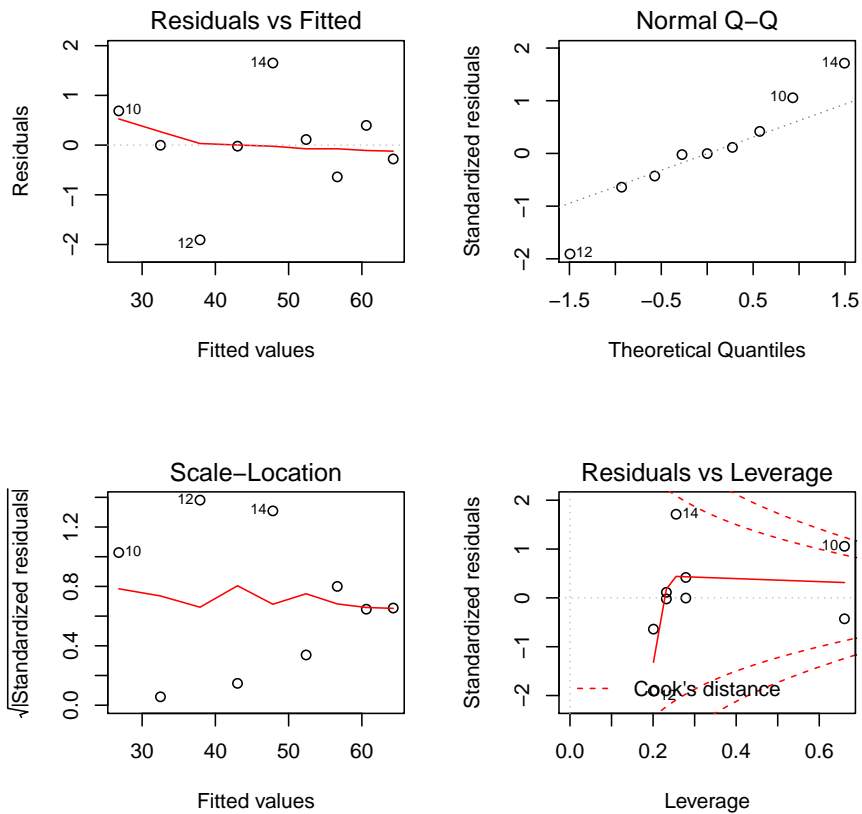
Residuals:
    Min       1Q   Median       3Q      Max
-1.90606 -0.27879 -0.00303  0.39697  1.65152

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.83333    1.42174   14.653 6.34e-06 ***
Time         6.12273    0.65281    9.379 8.34e-05 ***
I(Time^2)   -0.14394    0.06367   -2.261  0.0645 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.117 on 6 degrees of freedom
Multiple R-squared:  0.9944,    Adjusted R-squared:  0.9925
F-statistic: 529.6 on 2 and 6 DF,  p-value: 1.787e-07

> par(mfrow=c(2,2))
> plot(prase2.lm)

```



Úkol 3: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v $\text{T}_{\text{E}}\text{X}_{\text{u}}$) nebo můžete každé řešení označit číslem úkolu a vše odevzdat ve zvláštním dokumentu.

Úkol 4: Pomocí následujícího lineárního modelu rozhodněte, jestli obě prasata mohla být při narození stejně těžká a jestli obě prasata rostla stejně rychle. Které prase roste rychleji? Je pozorovaný rozdíl rychlosti růstu statisticky signifikantní? Při interpretaci parametrů zvažte i matici modelu získanou příkazem `model.matrix(prase.lm)`.

```
> prase.lm=lm(weight~Time+ID+Time:ID)
> summary(prase.lm)
```

```
Call:
lm(formula = weight ~ Time + ID + Time:ID)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-1.9611 -0.7222 -0.2583  0.8653  2.6111
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.5278     0.9794   18.92 2.28e-11 ***
Time          6.2167     0.1740   35.72 3.74e-15 ***
ID18         4.9444     1.3851    3.57 0.00308 **
Time:ID18    -1.5333     0.2461   -6.23 2.20e-05 ***
```

```

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.348 on 14 degrees of freedom
Multiple R-squared:  0.9931,    Adjusted R-squared:  0.9916
F-statistic: 672.8 on 3 and 14 DF,  p-value: 2.306e-15

> model.matrix(prase.lm)

  (Intercept) Time ID18 Time:ID18
1           1     1     0         0
2           1     2     0         0
3           1     3     0         0
4           1     4     0         0
5           1     5     0         0
6           1     6     0         0
7           1     7     0         0
8           1     8     0         0
9           1     9     0         0
10          1     1     1         1
11          1     2     1         2
12          1     3     1         3
13          1     4     1         4
14          1     5     1         5
15          1     6     1         6
16          1     7     1         7
17          1     8     1         8
18          1     9     1         9
attr("assign")
[1] 0 1 2 3
attr("contrasts")
attr("contrasts")$ID
[1] "contr.treatment"

> anova(prase.lm2<-lm(weight~Time),prase.lm)

Analysis of Variance Table

Model 1: weight ~ Time
Model 2: weight ~ Time + ID + Time:ID
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     16 129.325
2     14  25.444  2    103.88 28.578 1.141e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Úkol 4: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v $\text{T}_{\text{E}}\text{X}$) nebo můžete každé řešení označit číslem úkolu a vše odevzdat ve zvláštním dokumentu.

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

```
> rstudent(prase.lm)
```

1	2	3	4	5	6
-0.68670802	-1.82846056	1.56606790	0.46734058	0.68599434	0.92629773
7	8	9	10	11	12
-0.43224783	-0.21752034	-0.91404596	-0.60226517	-0.28266594	-1.27247303
13	14	15	16	17	18
0.61686482	2.36851953	0.72426149	-0.20176454	0.05082167	-1.60985564

Úkol 5: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v \TeX) 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(prase.lm)
```

1	2	3	4	5	6	7	8
0.3777778	0.2611111	0.1777778	0.1277778	0.1111111	0.1277778	0.1777778	0.2611111
9	10	11	12	13	14	15	16
0.3777778	0.3777778	0.2611111	0.1777778	0.1277778	0.1111111	0.1277778	0.1777778
17	18						
0.2611111	0.3777778						

```
> influence.measures(prase.lm)
```

Influence measures of

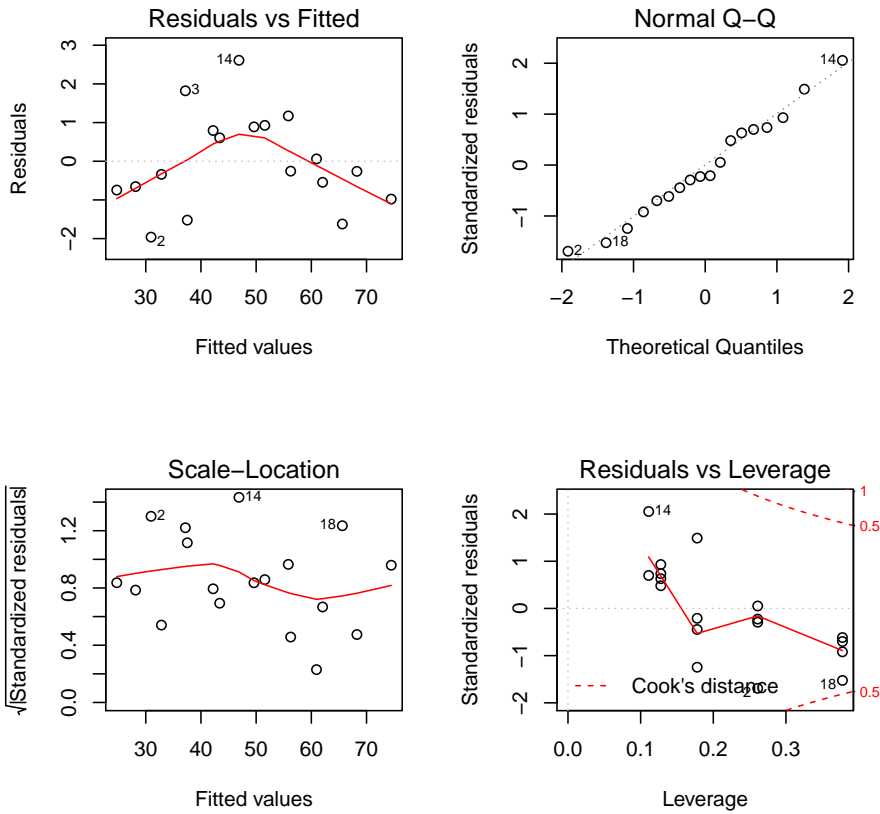
lm(formula = weight ~ Time + ID + Time:ID) :

	dfb.1_	dfb.Time	dfb.ID18	dfb.T.ID	dffit	cov.r	cook.d	hat	inf
1	-5.33e-01	4.50e-01	0.37660	-3.18e-01	-0.5351	1.875	0.074385	0.378	*
2	-1.06e+00	8.24e-01	0.74765	-5.83e-01	-1.0869	0.729	0.253015	0.261	*
3	6.60e-01	-4.46e-01	-0.46695	3.15e-01	0.7282	0.819	0.120109	0.178	
4	1.34e-01	-6.46e-02	-0.09471	4.57e-02	0.1789	1.443	0.008472	0.128	
5	1.11e-01	-8.85e-17	-0.07869	8.08e-17	0.2425	1.313	0.015284	0.111	
6	3.79e-02	1.28e-01	-0.02682	-9.05e-02	0.3545	1.194	0.031747	0.128	
7	3.65e-02	-1.23e-01	-0.02578	8.70e-02	-0.2010	1.545	0.010722	0.178	
8	4.84e-02	-9.80e-02	-0.03421	6.93e-02	-0.1293	1.794	0.004485	0.261	
9	3.54e-01	-5.98e-01	-0.25064	4.23e-01	-0.7122	1.685	0.128322	0.378	
10	5.31e-20	5.87e-17	-0.33029	2.79e-01	-0.4693	1.936	0.057682	0.378	*
11	-3.40e-18	3.29e-17	-0.11558	9.01e-02	-0.1680	1.776	0.007555	0.261	
12	1.65e-16	-8.49e-19	-0.37941	2.56e-01	-0.5917	1.023	0.083817	0.178	
13	-4.18e-17	2.97e-18	0.12501	-6.03e-02	0.2361	1.374	0.014582	0.128	
14	7.00e-17	-7.88e-17	0.27169	4.75e-17	0.8374	0.360	0.131883	0.111	
15	-3.00e-17	3.75e-17	0.02097	7.08e-02	0.2772	1.316	0.019887	0.128	
16	-8.72e-18	8.00e-18	0.01203	-4.06e-02	-0.0938	1.616	0.002362	0.178	
17	2.99e-18	-6.45e-19	-0.00799	1.62e-02	0.0302	1.819	0.000246	0.261	
18	-8.80e-17	3.56e-17	0.44143	-7.45e-01	-1.2544	1.045	0.353217	0.378	

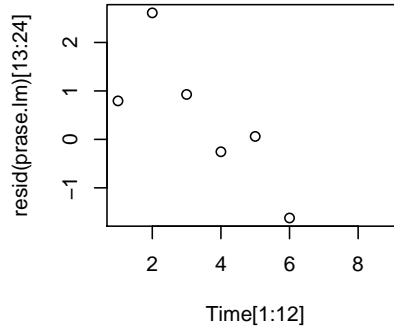
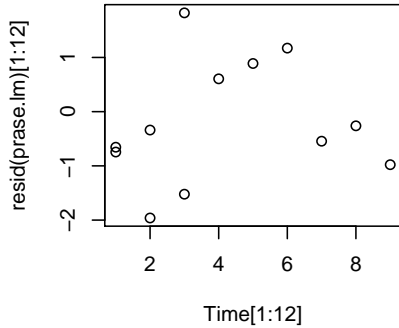
Úkol 6: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v \TeX) nebo můžete každé řešení označit číslem úkolu a vše odevzdat ve zvláštním dokumentu.

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

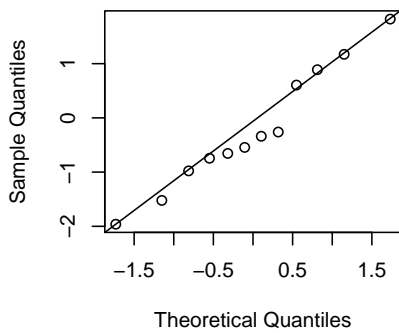

```
> par(mfrow=c(2,2))
> plot(prase.lm)
```



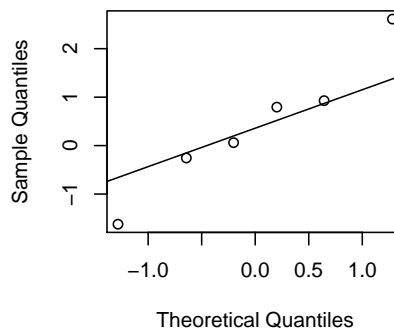
```
> par(mfrow=c(2,2))
> plot(resid(prase.lm)[1:12]~Time[1:12])
> plot(resid(prase.lm)[13:24]~Time[1:12])
> qqnorm(resid(prase.lm)[1:12])
> qqline(resid(prase.lm)[1:12])
> qqnorm(resid(prase.lm)[13:24])
> qqline(resid(prase.lm)[13:24])
```



Normal Q-Q Plot



Normal Q-Q Plot



Úkol 7: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v TeX u) 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 pro váhy prasat.

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

Shapiro-Wilk normality test

data:  resid(prase.lm)
W = 0.97366, p-value = 0.8629

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

Shapiro-Wilk normality test

data:  rstandard(prase.lm)
W = 0.97805, p-value = 0.9277

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

Shapiro-Wilk normality test

data:  rstudent(prase.lm)
W = 0.97904, p-value = 0.9395
```

Úkol 8: Svoji odpověď můžete buď vepsat přímo do tohoto boxu (v \TeX u) 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 pro váhy prasat.

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

Durbin-Watson test

```
data: prase.lm
DW = 1.5098, p-value = 0.0228
alternative hypothesis: true autocorrelation is greater than 0
```

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

Úkol 10: Jak byste (s využitím odhadů parametrů v lineárním modelu `prase.lm`) otestovali hypotézu, že obě prasata byla ve věku 7 týdnů stejně těžká?

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

Úkol 11: Stručně a výstižně shrňte výsledky získané analýzou lineárního modelu `prase.lm`, a vysvětlete, které předpoklady normálního lineárního modelu jsou v tomto případě porušeny a jak by bylo možné model `prase.lm` vylepšit.

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

Jenom pro kontrolu: datum narození je 19700131 a vybrali jsme prasata 12 a 18.