Lineární regrese

Komentované řešení pomocí programu R

Ústav matematiky
Fakulta chemicko inženýrská
Vysoká škola chemicko-technologická v Praze

Načtení vstupních dat

Vstupní data se nacházejí v souboru "data-lin_regrese.txt". Předpokládejme, že data jsou uložena na disku F ve složce Aplikovana_statistika.
Načtení vstupních dat do pracovního objektu DATA a vypsání na obrazovku provedeme příkazem

```
DATA<-read.csv("f:\\Aplikovana_statistika\\data-lin_regrese.txt", header=TRUE)
DATA
```

```
  koncentrace napetova_odezva
1          12           62.42
2          17           85.16
3          33          165.84
4          41          205.25
5          56          280.18
6          64          320.20
7          78          392.74
8          89          447.00
9          92          496.80
```

Základní ukazatele

Lineární regresní model **lregrese**, v němž je na napěťová odezva závislou proměnnou na vysvětlující proměnné koncentrace, zavedeme příkazem

```
lregrese<-lm(DATA$napetova_odezva~DATA$koncentrace)
```

Zdůrazněme, že jako první před vlnovkou píšeme do argumentu závislou proměnnou a za vlnovku vysvětlující proměnnou.

Nyní se budeme zajímat o základní ukazatele modelu $Y = \alpha + \beta x$. Tím prvním, co nás zajímá, jsou odhady koeficientů α, β. K tomu stačí nechat si vypsat objekt **lregrese**.

```
lregrese
Call:
lm(formula = DATA$napetova_odezva ~ DATA$koncentrace)

Coefficients:
     (Intercept)  DATA$koncentrace
          -5.723             5.201
```

Koeficient v prvním sloupci (Intercept) odpovídá odhadu parametru α a koeficient ve druhém sloupci je odhadem směrnice β. Většinu základních ukazatelů (včetně právě spočtených odhadů) lze jednoduše najednou vypsat pomocí příkazu **"summary"**.

```
summary(lregrese)

Call:
lm(formula = DATA$napetova_odezva ~ DATA$koncentrace)

Residuals:
    Min      1Q  Median      3Q     Max
-10.206  -6.970  -2.286   2.459  23.990

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -5.7233     7.9305  -0.722    0.494
DATA$koncentrace  5.2015     0.1312  39.658 1.69e-09 ***
--
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.04 on 7 degrees of freedom
Multiple R-squared:  0.9956,    Adjusted R-squared:  0.9949
F-statistic:  1573 on 1 and 7 DF,  p-value: 1.688e-09
```

Nyní si podrobně rozebereme, co nám program vypsal na konzoli

```
Residuals:
    Min      1Q  Median      3Q     Max
-10.206  -6.970  -2.286   2.459  23.990
```

. . . vypíše minimální, maximální reziduum a také kvartily reziduí.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -5.7233     7.9305  -0.722    0.494
DATA$koncentrace  5.2015     0.1312  39.658 1.69e-09 ***
--
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

. . . sloupec Estimate obsahuje odhady parametrů
. . . sloupec Std. Error jim odpovídající chyby
. . . sloupec t value příslušné t-statistiky
. . . sloupec Pr(>|t|) pravděpodobnost, že tyto statistiky padnou do kritického oboru (p-hodnoty)

```
Residual standard error: 11.04 on 7 degrees of freedom
Multiple R-squared:  0.9956,    Adjusted R-squared:  0.9949
F-statistic:  1573 on 1 and 7 DF,  p-value: 1.688e-09
```

. . . Residual standard error je chyba modelu Y.
. . . Multiple R-squared je tzv. koeficient vícenásobné determinace. Jedná se o vícenásobnou verzi klasického indexu determinace (ovšem my máme pouze jednoduchou regresi, takže oba koeficienty víceméně splývají). Koeficient determinace $\times 100$ popisuje, kolik procent rozptylu bylo modelem vysvětleno (tj. kolik procent dat bylo modelem dobře popsáno).
. . . Adjusted R-squared je obdobou koeficientu determinace, ale narozdíl od klasického zohledňuje počet parametrů modelu, takže budeme hledět na něj.
. . . F-statistic je příslušná statistika včetně stupňů volnosti a p-hodnoty.

Rezidua

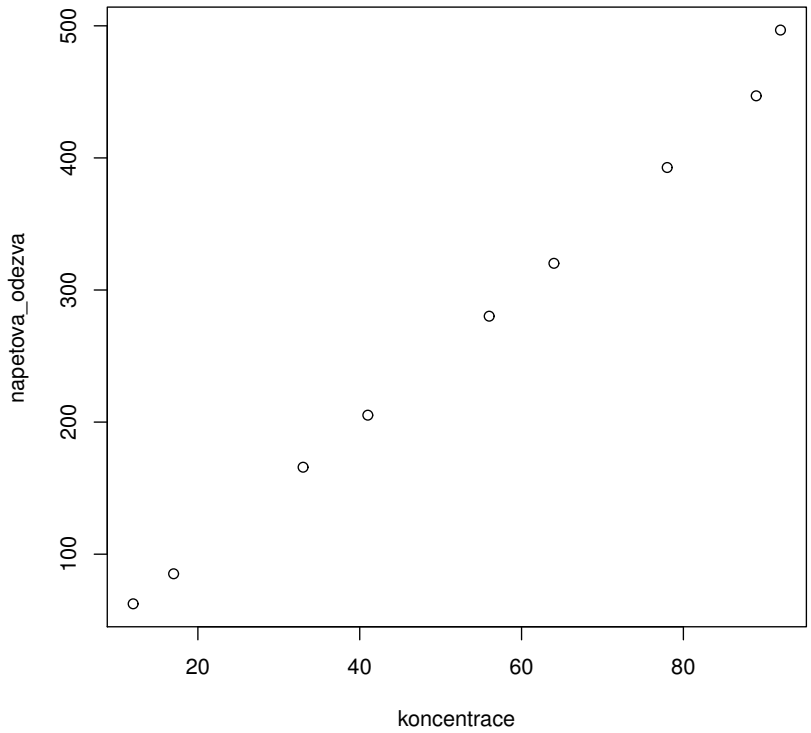
Už jsme si ukázali, jak vypsat některá rezidua pro náš model. Rezidua představují odhad nepozorovatelných náhodných chyb v měření. Chceme-li vypsat všechna rezidua použijeme příkaz

```
residuals(lregrese)
          1           2           3           4           5           6
 5.72585188  2.45859697 -0.08461872 -2.28622657 -5.37799129 -6.96959913
          7           8           9
-7.24991287 -10.20587366 23.98977340
```

Ověření předpokladů - linearita

Nejprve graficky ověříme, že lineární model je pro daná data vhodný - zkonstruujeme bodový graf závislosti napěťové odezvy na koncentraci. Základní příkaz pro nakreslení (nejen) bodového grafu

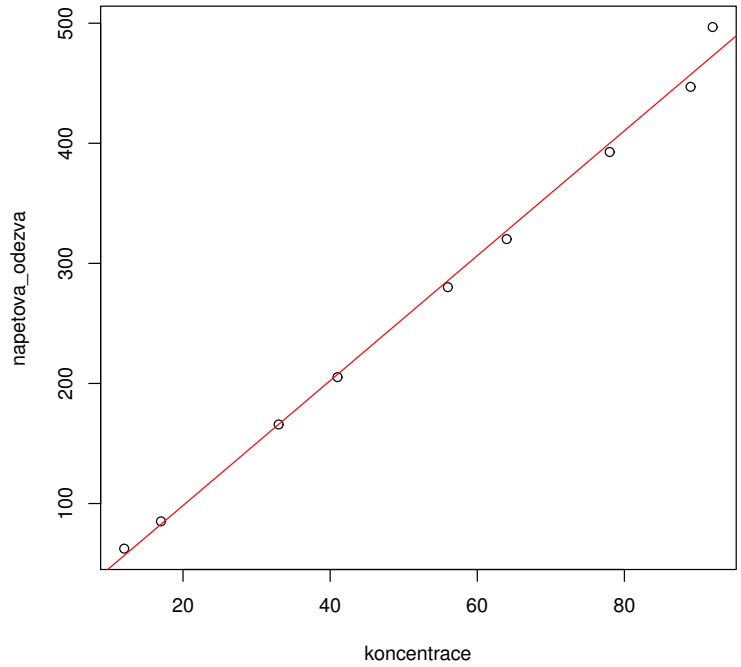
```
plot(DATA)
```



Zdůrazněme, že tento příkaz použijeme na původní data.

Pokud budeme chtít do grafu vykreslit i regresní přímku, můžeme použít příkaz

```
> abline(lregrese,col="red")
```

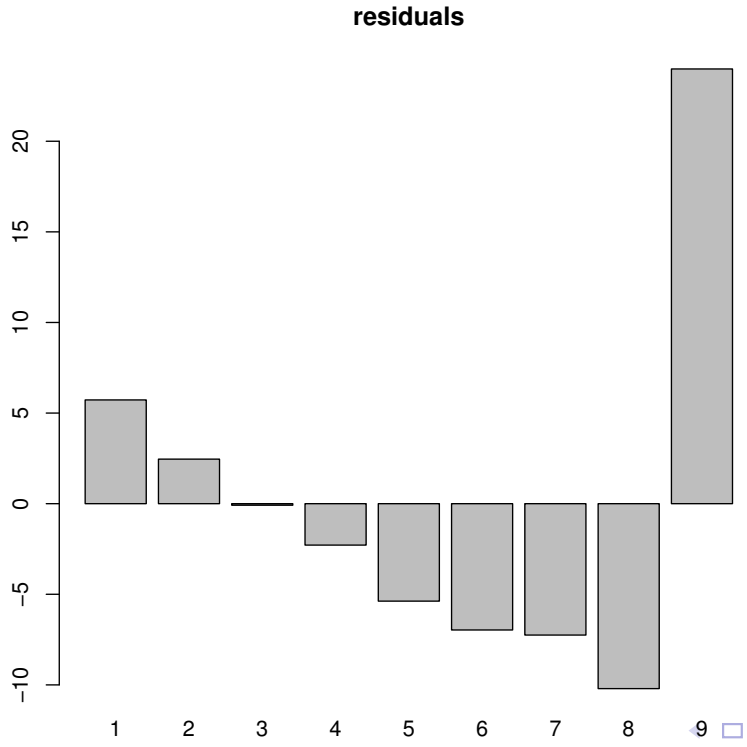


Graf skutečně podporuje naší hypotézu o lineární závislosti. Pouze poslední bod se výrazněji odchyluje od regresní přímky. Detailnější pohled přinese analýza reziduí.

Ověření předpokladů - rezidua

Abychom posoudili vzájemnou nekorelovanost i neměnný rozptyl náhodných odchylek, vykreslíme si graf reziduí. Zkonstruujeme sloupcový graf, kde každý sloupec bude reprezentovat jedno měření:

```
> barplot(residuals(lregrese), main="residuals")
```



Opravená data

Graf reziduí jasně ukazuje, jak moc se poslední pozorování vymyká, proto ho vyloučíme z našeho modelu. Z dat ho můžeme odebrat jednouduchým příkazem

```
DATA<-DATA[-9,]
DATA
  koncentrace napetova_odezva
1          12           62.42
2          17           85.16
3          33          165.84
4          41          205.25
5          56          280.18
6          64          320.20
7          78          392.74
8          89          447.00
```

Regresní přímku bychom mohli uložit do stejného objektu **lregrese**, ale obecně je potřeba dát si na přepisování (stejně jako v každém programovacím jazyce) pozor, aby nám někde nezůstala uložená nějaká stará data. Pokud chceme mít jistotu, můžeme vymazat všechno pomocí příkazu

```
rm(list=ls(all=TRUE))
```

Tento příkaz ovšem vymaže úplně všechno, takže musíme znovu načíst data a znovu z nich odebrat poslední pozorování

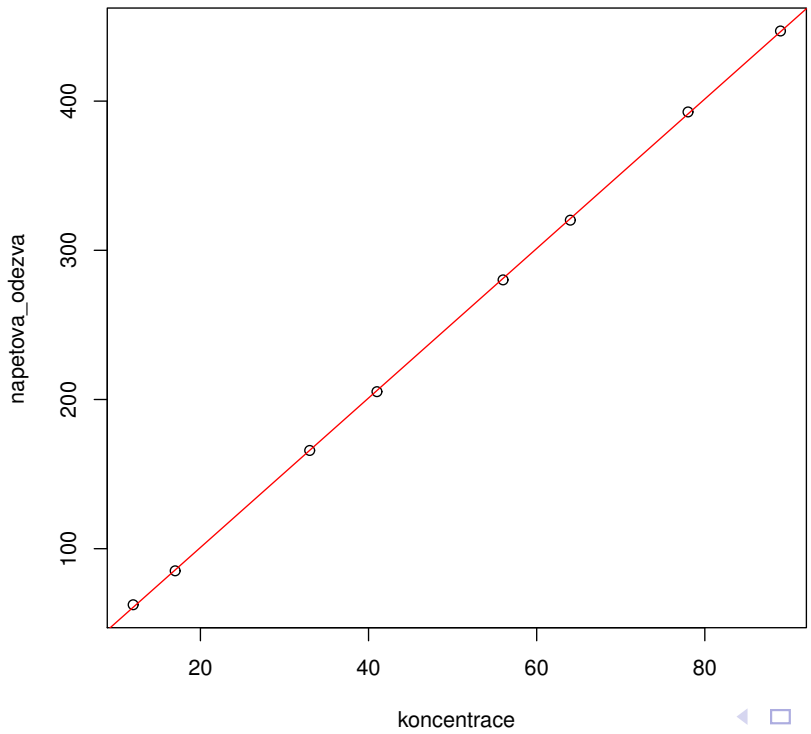
```
DATA<-read.table("C:\Aplikovana_statistika\data-lin_regrese.txt", header=TRUE)
DATA<-DATA[-9,]

lregrese<-lm(DATA$napetova_odezva~DATA$koncentrace)
```

Příkazem **summary** si opět analogicky vypíšeme většinu potřebných údajů

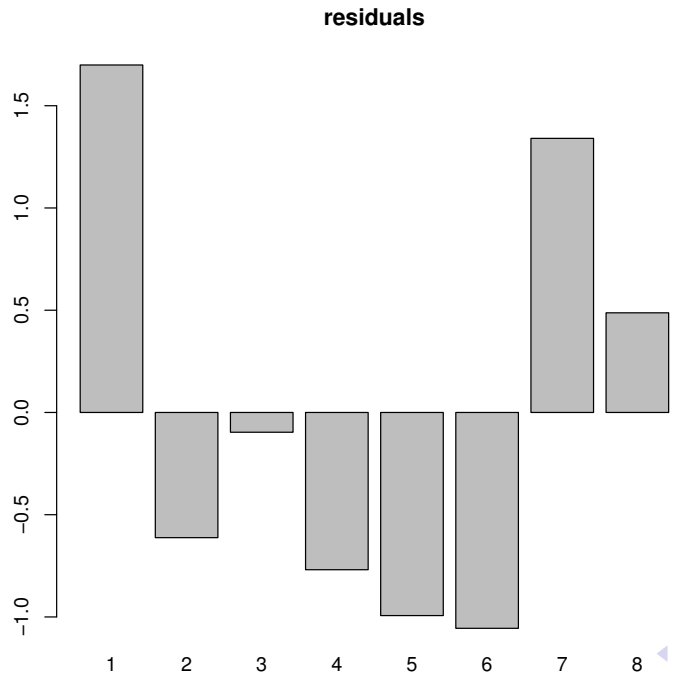
```
summary(lregrese)
Call:
lm(formula = DATA$napetova_odezva ~ DATA$koncentrace)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0556 -0.8251 -0.3545  0.7006  1.6992

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.59740    0.86621    0.69    0.516
DATA$koncentrace  5.01028    0.01567  319.70 6.32e-14 ***
--
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.155 on 6 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 1.022e+05 on 1 and 6 DF,  p-value: 6.321e-14
```

Jak vidíme, residual standard error se výrazně zmenšila, naproti tomu statistika F je výrazně vyšší, což znamená těsnější proložení regresní přímky. Také směrodatné odchylky odhadů se snížily a index determinace je bližší 1.

```
plot(DATA)
abline(lregrese,col="red")
```



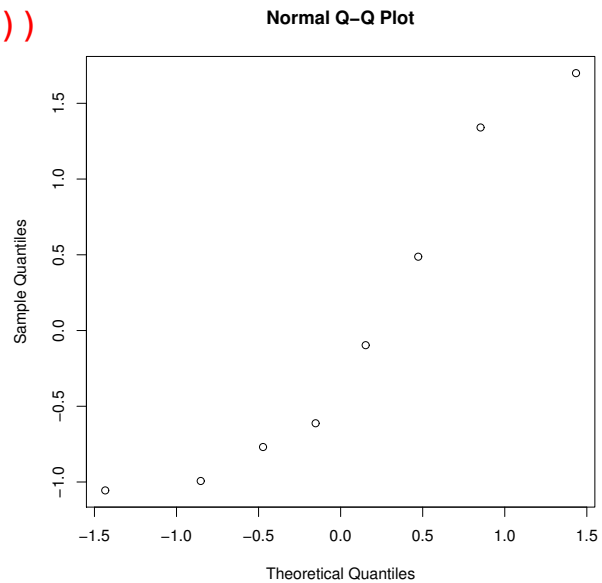
```
residuals(lregrese)
 1           2           3           4           5           6
 1.69919438 -0.61222662 -0.09677384 -0.76904744 -0.99331046 -1.05558406
           7           8
 1.34043713  0.48731092
barplot(residuals(lregrese), main="residuals")
```



Ověření předpokladů - normalita

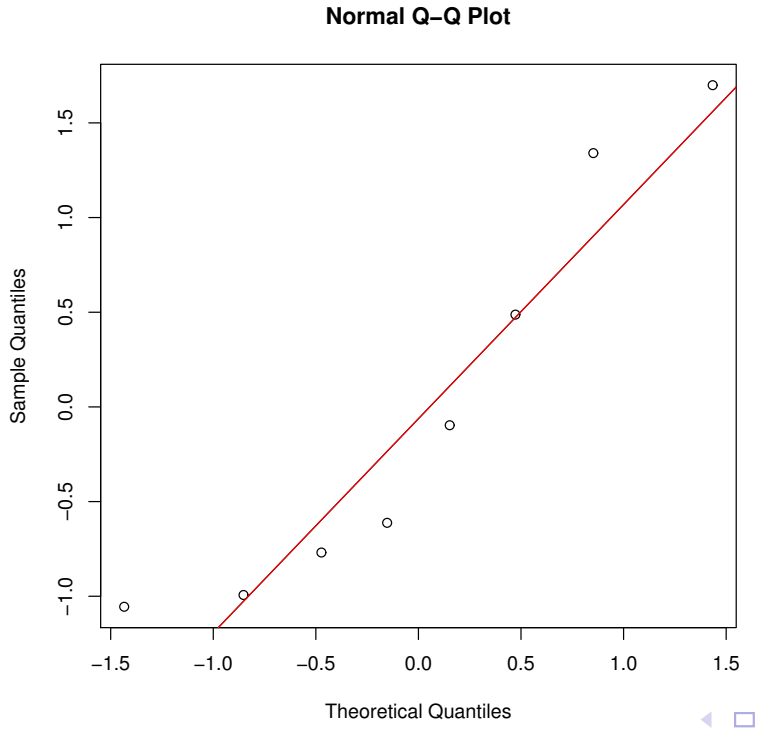
Z grafu vidíme, že rezidua už vypadají náhodně a nesystematicky, což je potřebný předpoklad pro základní model lineární regrese. Z grafu dat zase vidíme, že data jsou zjevně přímkou dobře nafitovaná. K ověření dalšího předpokladu - normality - si vykreslíme Q-Q graf, neboť na histogram máme málo pozorování.

```
qqnorm(residuals(lregrese))
```



Q-Q graf, který jsme zkonstruovali, znázorňuje vztah mezi rezidui a teoretickými kvantily normálního rozdělení. Normalitu dat by indikovalo, kdyby se body Q-Q grafu soustředily kolem přímky $y = x$. Pro větší názornost si tuto přímku do obrázku můžeme vykreslit (uděláme to červenou barvou).

```
qqline(residuals(lregrese),col="red")
```



Mezi body Q-Q grafu a přímkou $y = x$ jsou sice patrné rozdíly, to ovšem vzhledem k malému množství dat ještě nemusí znamenat, že předpoklad normality chyb není splněn. Statistickou významnost normality reziduí bychom mohli otestovat některým z testů normality (existují testy, které fungují relativně dobře i pro malý rozsah výběru). To je ovšem nad rámec tohoto cvičení, a proto budeme nadále normalitu předpokládat.

Úloha (A)

Zda kalibrační křivka prochází počátkem je úloha ekvivalentní testu hypotézy o nulovosti koeficientu posunu (parametr α), kterou otestujeme pomocí oboustranného intervalu spolehlivosti. Příkaz **confint** nám (pokud nespecifikujeme blíže, které) vypíše oboustranné intervaly spolehlivosti pro jednotlivé parametry. Argumentem level=0.95 udáváme, že chceme 95% interval spolehlivosti.

```
confint(lregrese, level=0.95)
                    2.5 %   97.5 %
(Intercept)      -1.522139 2.716929
DATA$koncentrace  4.971937 5.048632
```

Nás momentálně zajímá řádek (Intercept), neboť právě to je odhad koeficientu posunu. Jak vidíme, interval spolehlivosti pro tento koeficient pokrývá nulu, nemůžeme tedy na hladině významnosti 95% zamítnout hypotézu o jeho nulovosti.

Úloha (E)

Obdobně jako v úloze (A) nás zajímá, jestli směrnice regresní přímky (parametr β) nemůže být nulový. Máme zkonstruovaný interval spolehlivosti i pro tento parametr (viz předchozí slide)

```
confint(lregrese, level=0.95)
                    2.5 %   97.5 %
(Intercept)      -1.522139 2.716929
DATA$koncentrace  4.971937 5.048632
```

Jak vidíme, interval spolehlivosti pro směrnici regresní přímky je velmi úzký a neobsahuje nulu. Tvrzení o nulové směrnici tedy můžeme (s 95% spolehlivostí) zamítnout.

Úloha (C)

Nyní si určíme predikční pás pro vypočtenou regresní přímku, opět při hladině 95%. V programu **R** pro spočtení predikčního intervalu (uložíme si ho do proměnné **predikcni_int**) existuje opět příkaz

```
predikcni_int<-predict(lregrese, interval="prediction", level=0.95)
```

Vypíšeme si tento predikční interval

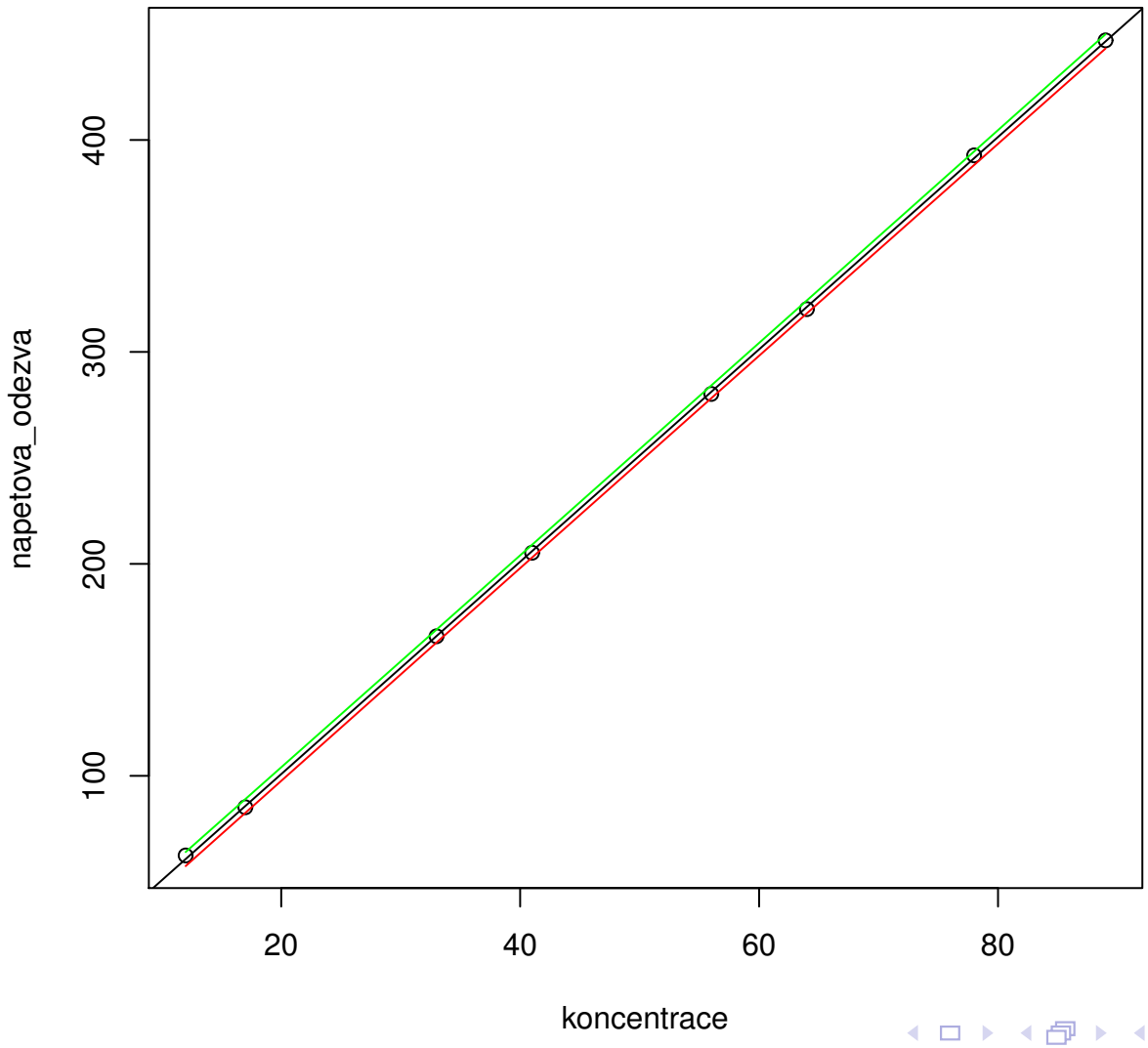
```
predikcni_int
        fit       lwr       upr
1  60.72081  57.40948  64.03213
2  85.77223  82.53784  89.00661
3 165.93677 162.88003 168.99351
4 206.01905 203.00787 209.03023
5 281.17331 278.16396 284.18266
6 321.25558 318.20258 324.30859
7 391.39956 388.20003 394.59909
8 446.51269 443.14205 449.88333
```

Ve sloupci "fit" máme odhad hodnoty $Y(x)$, ve sloupci "lwr" dolní mez intervalového odhadu a ve sloupci "upr" horní mez intervalového odhadu.

Chceme-li si do jednoho obrázku vykreslit pozorování, regresní přímku a interval spolehlivosti budeme muset zkombinovat několik příkazů

```
plot(DATA)
abline(lregrese)
lines(DATA[,1],predikcni_int[,2],type="l",col="red")
lines(DATA[,1],predikcni_int[,3],type="l",col="green")
```

První příkaz vykreslí data, druhý regresní přímku (černá), třetí a čtvrtý vykreslí spojnice mezi jednotlivými dolními (červená), respektive horními (zelená) hodnotami predikčního intervalu.


napetova_odezva
100
200
300
400
20
40
60
80
koncentrace

Interpretace - úloha (B)

Vzhledem k předchozí analýze reziduí lze usuzovat, že poslední měření z původních dat je systematicky chybné. Po odebrání tohoto měření jsou už všechna pozorování uvnitř 95% predikčního pásu z úlohy (C) a rezidua nevykazují žádné výrazně odlehlé hodnoty, takže další systematicky chybné měření už nepředpokládáme.

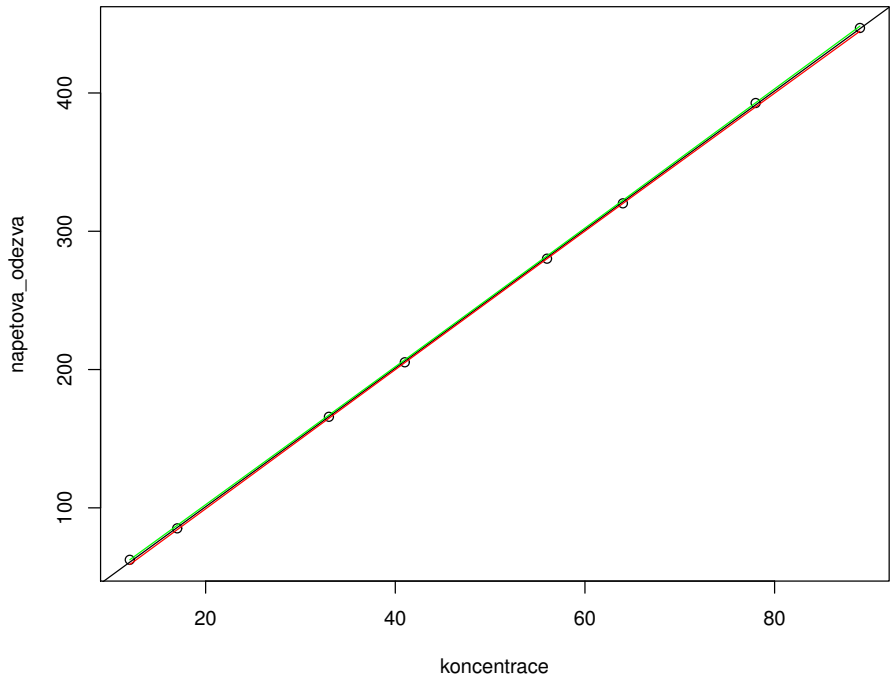
Chceme-li si spočítat 95 % pás spolehlivosti, použijeme analogický příkaz

```
pas_spolehlivosti<-predict(lregrese, interval="confidence", level=0.95)
pas_spolehlivosti
```

```
        fit       lwr       upr
1  60.72081  58.99347  62.44814
2  85.77223  84.19742  87.34704
3 165.93677 164.76954 167.10401
4 206.01905 204.97695 207.06115
5 281.17331 280.13651 282.21011
6 321.25558 320.09816 322.41301
7 391.39956 389.89764 392.90149
8 446.51269 444.67422 448.35116
```

Vykreslení pásu spolehlivosti probíhá obdobným způsobem jako pro pás predikční

```
plot(DATA)
abline(lregrese)
lines(DATA[,1],pas_spolehlivosti[,2],type="l",col="red")
lines(DATA[,1],pas_spolehlivosti[,3],type="l",col="green")
```


napetova_odezva
100
200
300
400
20
40
60
80
koncentrace

Jak vidíme, pás spolehlivosti je velice úzký, pojďme se tedy zaměřit jen na okolí bodu 50. Nakreslíme si zvětšený graf na okolí bodu 50. To se provede pomocí argumentů xlim, ylim.

```
plot(DATA,xlim=c(20,80),ylim=c(200,300))
abline(lregrese)
lines(DATA[,1],pas_spolehlivosti[,2],type="l",col="red")
lines(DATA[,1],pas_spolehlivosti[,3],type="l",col="green")
```

Dále do grafu naneseme úsečku vyjadřující požadovanou přesnost. Souřadnice jejího počátečního koncového bodu vypočteme následovně:

1. Počáteční bod: $x-$ová souřadnice: $x1 = 50 - 0,2 * 50$
2. Koncový bod: $x-$ová souřadnice: $x2 = 50 + 0,2 * 50$
3. $y-$ová souřadnice obou bodů je funkční hodnotou odhadnuté regresní přímky v bodě 50.

Celé to v programu **R** proběhne následovně (do vektoru koeficienty si uložíme spočtené odhady koeficientů regresní přímky)

```
koeficienty<-coefficients(lregrese)   ##vektor obsahujici odhady koeficientu

y<-koeficienty[1]+koeficienty[2]*50      ## vypocet y(50) na regresni primce
x1<-50-0.02*50
x2<-50+0.02*50
lines(c(x1,x2),c(y,y),type="l")
```

Příkaz lines dokreslí do stavajícího grafu úsečku, jejíž krajní body mají souřadnice $(x1, y)$ a $(x2, y)$.

Interpretace: 95% pás spolehlivosti obsahuje skutečnou (neznámou) regresní funkci s pravděpodobností 95 % a tento pás je výrazně užší než požadovaná přesnost 2 %. Lze tedy tvrdit, že na hladině významnosti 95 % lze určit koncentraci s přesností lepší než 2 %.