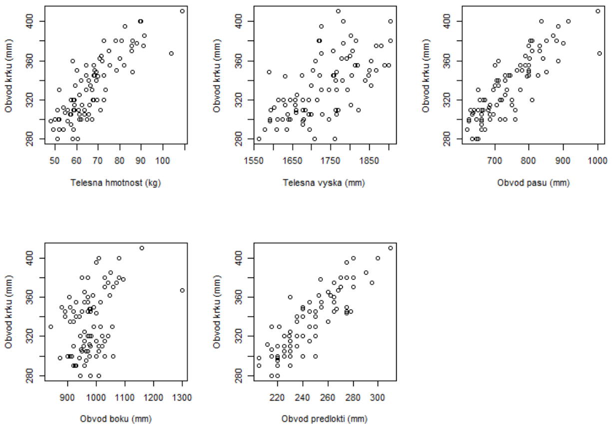
Vícenásobná lineární regrese – vzorový příklad

V souboru cneck.txt máme k dispozici antropometrická data mladých dospělých lidí (převážně studentů vysokých škol z Brna a Ostravy). Chceme modelovat závislost obvodu krku (proměnná neck.C) na tělesné hmotnosti (proměnná body.W), tělesné výšce (proměnná body.H), obvodu pasu (proměnná waist.C), obvodu boků (proměnná hip.C) a obvodu předloktí (proměnná antb.C). Hmotnost byla měřena v kg, délkové míry v mm.
Načteme data a podíváme se na ně. Soubor neobsahuje žádná chybějící pozorování.

```
> neck <- read.table("cneck.txt",header=T)
> summary(neck)
       id          sex       body.W        body.H         waist.C          hip.C           antb.C          neck.C
 Min.   :  1.00   f:38   64,5   : 4   Min.   :1563   Min.   : 620.0   Min.   : 840.0   Min.   :205.0   Min.   :280.0
 1st Qu.: 44.00   m:49   68     : 4   1st Qu.:1660   1st Qu.: 663.5   1st Qu.: 945.0   1st Qu.:225.0   1st Qu.:306.0
 Median : 91.00          52     : 3   Median :1725   Median : 730.0   Median : 970.0   Median :240.0   Median :330.0
 Mean   : 92.23          57,5   : 3   Mean   :1729   Mean   : 740.1   Mean   : 979.9   Mean   :244.5   Mean   :332.9
 3rd Qu.:139.50          58,5   : 3   3rd Qu.:1792   3rd Qu.: 800.0   3rd Qu.:1010.0   3rd Qu.:263.5   3rd Qu.:355.0
 Max.   :188.00          59     : 3   Max.   :1906   Max.   :1005.0   Max.   :1300.0   Max.   :310.0   Max.   :410.0
```

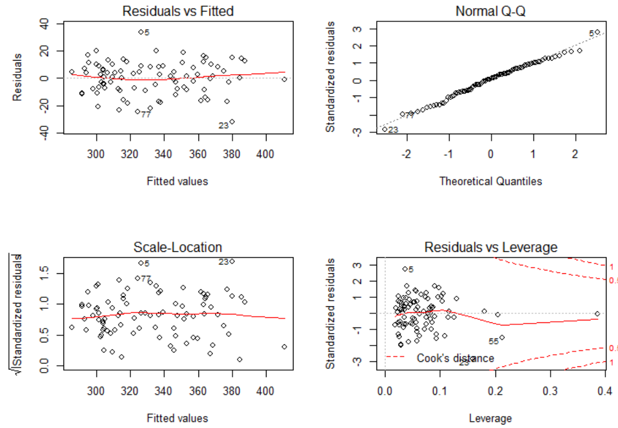
Vykreslíme si bodové diagramy pro dvojice (obvod krku, hmotnost); (obvod krku, výška); (obvod krku, obvod pasu); (obvod krku, obvod boků) a (obvod krku, obvod předloktí).

```
> par(mfrow=c(2,3))
> plot(neck$body.W, neck$neck.C, xlab='Telesna hmotnost (kg)', ylab='Obvod krku (mm)')
> plot(neck$body.H, neck$neck.C, xlab='Telesna vyska (mm)', ylab='Obvod krku (mm)')
> plot(neck$waist.C, neck$neck.C, xlab='Obvod pasu (mm)', ylab='Obvod krku (mm)')
> plot(neck$hip.C, neck$neck.C, xlab='Obvod boku (mm)', ylab='Obvod krku (mm)')
> plot(neck$antb.C, neck$neck.C, xlab='Obvod predlokti (mm)', ylab='Obvod krku (mm)')
```



Bodové diagramy naznačují, že je mezi dvojicemi lineární závislost. Sestavíme regresní model a pomocí analýzy reziduí ověříme předpoklady modelu.

```
> model1 <- lm(neck.C ~ body.W + body.H + waist.C + hip.C + antb.C, data=neck)
> par(mfrow=c(2,2))
> plot(model1)
```



Interpretace grafu je stejná jako u jednoduchého regresního modelu. Ověříme předpoklady i pomocí vhodných testů. Pomocí t-testu otestujeme hypotézu, že rezidua mají nulovou střední hodnotu. Normalitu reziduí ověříme pomocí Shapirova-Wilkova testu a nezávislost reziduí ověříme pomocí Durbinova-Watsonova testu (z knihovny car).

```
> t.test(model1$residuals)
        One Sample t-test
data:  model1$residuals
t = 9.3512e-17, df = 86, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -2.555173  2.555173
sample estimates:
   mean of x
1.201944e-16

> shapiro.test(model1$residuals)
        Shapiro-Wilk normality test
data:  model1$residuals
W = 0.9902, p-value = 0.7645

> library(car)
> durbinWatsonTest(model1)
 lag Autocorrelation D-W Statistic p-value
   1       0.1541587      1.678153    0.12
 Alternative hypothesis: rho != 0
```

Hypotézu o nulové střední hodnotě reziduí ...., protože t-test nabývá hodnoty .... s p-hodnotou ...., z grafického posouzení také nevidíme problém.

Shapirův-Wilkuv test nabývá hodnoty .... s p-hodnotou ...., v kvantil-kvantilovém grafu jsou rezidua ...., předpoklad normality tedy považujeme za ....

Předpoklad rovnosti rozptylů se na základě grafického posouzení zdá ....

Durbin-Watsonův test nabývá hodnoty .... s p-hodnotou ...., tedy .... nezávislost reziduí.
Předpoklady modelu jsou tedy ....

Podívejme se, jestli v našem modelu není problém s multikolinearitou. Vypočítáme si korelační koeficienty mezi nezávislými proměnnými a také hodnoty koeficientu V IF pro proměnné sestaveného modelu.

```
> cor(neck[,c('body.W', 'body.H', 'waist.C', 'hip.C', 'antb.C')])
           body.W    body.H   waist.C     hip.C    antb.C
body.W  1.0000000 0.6086383 0.9047087 0.7604090 0.8810742
body.H  0.6086383 1.0000000 0.4591687 0.2303759 0.5851208
waist.C 0.9047087 0.4591687 1.0000000 0.6539080 0.8520787
hip.C   0.7604090 0.2303759 0.6539080 1.0000000 0.5251877
antb.C  0.8810742 0.5851208 0.8520787 0.5251877 1.0000000

> vif(model1)
   body.W    body.H   waist.C     hip.C    antb.C
18.895276  2.307445  6.812388  3.904779  6.116750
```

Vidíme, že jak korelační koeficienty, tak koeficienty V IF nabývají vysokých hodnot, lze tedy soudit na existenci multikolinearity. Vypíšeme si podrobné informace o modelu:

```
> summary(model1)
Call:
lm(formula = neck.C ~ body.W + body.H + waist.C + hip.C + antb.C,
    data = neck)

Residuals:
    Min      1Q  Median      3Q     Max
-32.266  -8.030   1.169   8.493  33.577

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 165.63910   64.79600   2.556   0.0124 *
body.W        1.02594    0.46867   2.189   0.0315 *
body.H        0.04039    0.02314   1.745   0.0847 .
waist.C       0.18260    0.04025   4.537 1.96e-05 ***
hip.C        -0.18166    0.04070  -4.463 2.58e-05 ***
antb.C        0.29120    0.14144   2.059   0.0427 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.35 on 81 degrees of freedom
Multiple R-squared:  0.8573,   Adjusted R-squared:  0.8485
F-statistic: 97.33 on 5 and 81 DF,  p-value: < 2.2e-16
```

MNČ odhady regresních koeficientů a jejich interpretace:
β_0 = ...
β_1 = ...
β_2 = ...
β_3 = ...
β_4 = ...
β_5 = ...
Odhadnutá regresní funkce má tvar ....
Index determinace ID2 = ...
Adjustovaný index determinace ID2adj = ...
Celkový F-test na hladině významnosti 0:05:
F = ...
p-hodnota = ...
závěr ....
Dílčí t-testy

parametr	hodnota testovací statistiky	p-hodnota	závěr
β_0			
β_1			
β_2			
β_3			
β_4			
β_5			

Intervaly spolehlivosti pro regresní koeficienty:

```
> confint(model1)
                   2.5 %       97.5 %
(Intercept) 36.715381614 294.56281096
body.W       0.093443025   1.95844313
body.H      -0.005657005   0.08643304
waist.C      0.102513283   0.26268666
hip.C       -0.262652856  -0.10067593
antb.C       0.009770142   0.57262713
```

Z výsledku dílcích testu vidíme, že proměnná body.H není na hladině 0,05 významná, sestavíme model, který ji neobsahuje.

```
> model2 <- lm(neck.C ~ body.W + waist.C + hip.C + antb.C, data=neck)
> summary(model2)

Call:
lm(formula = neck.C ~ body.W + waist.C + hip.C + antb.C, data = neck)

Residuals:
    Min      1Q  Median      3Q     Max
-36.799  -7.585  -0.460   8.523  33.903

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 255.80441   39.59226   6.461 6.96e-09 ***
body.W        1.49670    0.38801   3.857 0.000227 ***
waist.C       0.15765    0.03809   4.139 8.42e-05 ***
hip.C        -0.21493    0.03641  -5.903 7.74e-08 ***
antb.C        0.28727    0.14318   2.006 0.048114 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.51 on 82 degrees of freedom
Multiple R-squared:  0.8519,   Adjusted R-squared:  0.8447
F-statistic:   118 on 4 and 82 DF,  p-value: < 2.2e-16
```

Z podrobných informací o druhém modelu vidíme, že adjustovaný index determinace je o něco nižší, zřejmě tedy i proměnná body.H přispívá k vysvětlení variability obvodu krku.