4 Mnohonásobný lineární regresní model

Příklad 1. 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 kilogramech, délkové míry v milimetrech.

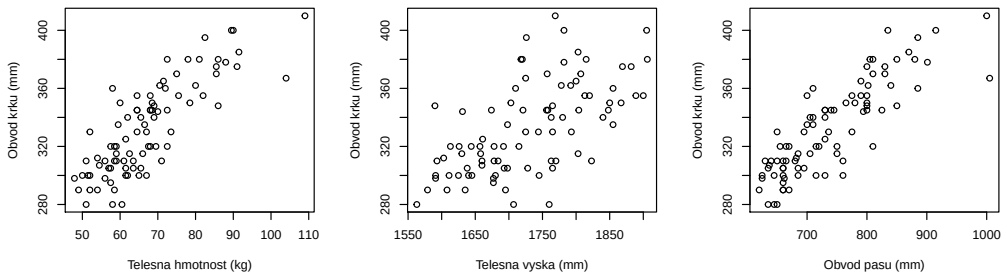
Načteme data a podíváme se na ně. Soubor neobsahuje žádná chybějící pozorování.

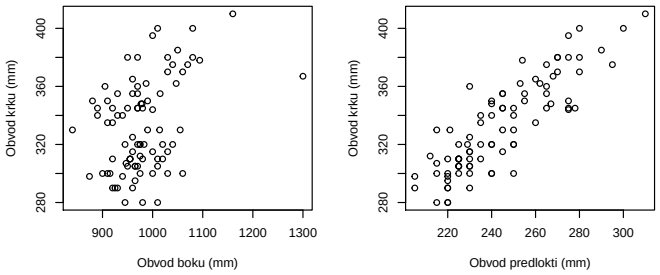
```
neck <- read.table("DATA/cneck.txt",header=T)
summary(neck)
```

```
##        id          sex        body.W           body.H        waist.C
##  Min.   :  1.00   f:49   Min.   : 47.90   Min.   :1563   Min.   : 620.0
##  1st Qu.: 44.00   m:38   1st Qu.: 58.50   1st Qu.:1660   1st Qu.: 663.5
##  Median : 91.00          Median : 65.50   Median :1725   Median : 730.0
##  Mean   : 92.23          Mean   : 67.32   Mean   :1729   Mean   : 740.1
##  3rd Qu.:139.50          3rd Qu.: 72.50   3rd Qu.:1792   3rd Qu.: 800.0
##  Max.   :188.00          Max.   :109.00   Max.   :1906   Max.   :1005.0
##      hip.C            antb.C          neck.C
##  Min.   : 840.0   Min.   :205.0   Min.   :280.0
##  1st Qu.: 945.0   1st Qu.:225.0   1st Qu.:306.0
##  Median : 970.0   Median :240.0   Median :330.0
##  Mean   : 979.9   Mean   :244.5   Mean   :332.9
##  3rd Qu.:1010.0   3rd Qu.:263.5   3rd Qu.:355.0
##  Max.   :1300.0   Max.   :310.0   Max.   :410.0
```

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

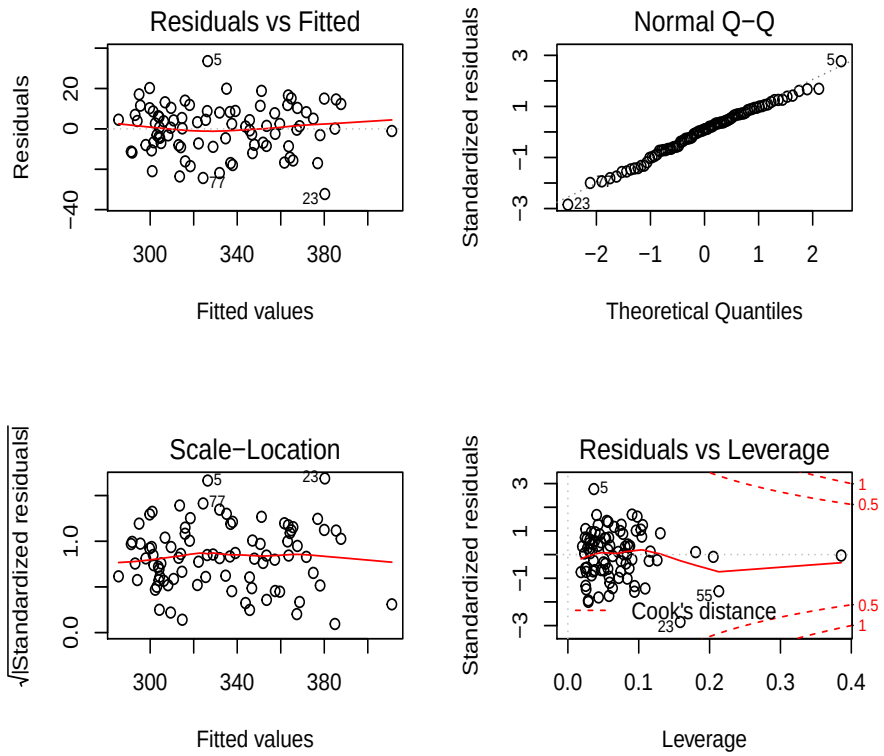
```
par(mfrow=c(2,3))
plot(neck$body.W, neck$neck.C, xlab='Telesna hmotnost (kg)', ylab='Obvod krku (mm)')
plot(neck$body.H, neck$neck.C, xlab='Telesna vyska (mm)', ylab='Obvod krku (mm)')
plot(neck$waist.C, neck$neck.C, xlab='Obvod pasu (mm)', ylab='Obvod krku (mm)')
plot(neck$hip.C, neck$neck.C, xlab='Obvod 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 grafů je stejná jako u jednoduchého regresního modelu. Ověřme 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)

## Loading required package: carData

durbinWatsonTest(model1)

##  lag Autocorrelation D-W Statistic p-value
##    1       0.1541587      1.678153    0.15
##  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-Wilkův 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 VIF 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 VIF nabývají vysokých hodnot, lze tedy soudit na existenci multikolinearity. Vypišme 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 koeficientů a jejich interpretace:
$\beta_0 =$ .....
$\beta_1 =$ .....
$\beta_2 =$ .....
$\beta_3 =$ .....
$\beta_4 =$ .....
$\beta_5 =$ .....

Odhadnutá regresní funkce má tvar .....

Index determinace (někdy nazýván koeficient determinace a značem R^2 místo ID^2):
$ID^2 =$ .....
Adjustovaný index determinace (někdy nazýván adjustovaný koeficient determinace a značem R^2_{adj} místo ID^2_{adj}):

ID^2_{adj} = ...

Celkový F-test na hladině významnosti 0.05:
F = ...
p-hodnota = ...
závěr .....

Dílčí t-testy
β_0

- hodnota testovací statistiky .....
- p-hodnota .....
- závěr .....

β_1

- hodnota testovací statistiky .....
- p-hodnota .....
- závěr .....

β_2

- hodnota testovací statistiky .....
- p-hodnota .....
- závěr .....

β_3

- hodnota testovací statistiky .....
- p-hodnota .....
- závěr .....

β_4

- hodnota testovací statistiky .....
- p-hodnota .....
- závěr .....

β_5

- hodnota testovací statistiky .....
- p-hodnota .....
- závěr .....

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
```

Interval spolehlivosti pro β_0: .....
Interval spolehlivosti pro β_1: .....
Interval spolehlivosti pro β_2: .....
Interval spolehlivosti pro β_3: .....
Interval spolehlivosti pro β_4: .....
Interval spolehlivosti pro β_5: .....

Z výsledků dílčích testů 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 druhého modelu vidíme, že adjustovaný index determinace je nižší, zřejmě tedy i proměnná body.H přispívá k vysvětlení variability obvodu krku.

Zkusme odhalit nejlepší model pomocí STEPWISE procedury, k tomu v R slouží funkce step(). Začněme s metodou backward. Jako vstupní argument je maximální model, který chceme uvažovat, a upřesnění směru, tedy direction='backward'. Metoda v prvním kroku vypouští vždy jen jednu proměnnou a zjišťuje, jestli se vypuštěním konkrétní jedné proměnné model zlepší. Poté vypustí tu, která vede k největšímu zlepšení modelu. V dalším kroku pracuje s modelem, který ji neobsahuje, a u něj opět vypouští jednotlivé proměnné. Opět vybere tu, jejíž vypuštění vede k největšímu zlepšení modelu, a takto pokračuje dále. Pokud metoda během kroku zjistí, že vypuštění jakékoli proměnné nevede k zlepšení, proces končí.

```
model.back <- step(lm(neck.C ~ body.W + body.H + waist.C + hip.C + antb.C, data=neck),
                   direction='backward')

## Start:  AIC=443.21
## neck.C ~ body.W + body.H + waist.C + hip.C + antb.C
##
##           Df Sum of Sq   RSS    AIC
## <none>                 12361 443.21
## - body.H   1    464.81 12826 444.42
## - antb.C   1    646.82 13008 445.64
## - body.W   1    731.29 13092 446.21
## - hip.C    1   3039.71 15401 460.33
## - waist.C  1   3140.65 15502 460.90

model.back

##
## Call:
## lm(formula = neck.C ~ body.W + body.H + waist.C + hip.C + antb.C,
##     data = neck)
##
## Coefficients:
## (Intercept)       body.W       body.H      waist.C        hip.C
##   165.63910      1.02594      0.04039      0.18260     -0.18166
##      antb.C
##     0.29120
```

Vidíme, že pro náš model metoda skončila hned v prvním kroku, tedy vypuštění jakékoli proměnné nevede k lepšímu modelu. Informace o výsledném modelu si můžeme vypsat pomocí funkce summary(), jak jsme zvyklí.

```
summary(model.back)

##
## 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
```

Podívejme se nyní na metodu forward. Jako vstupní argument je minimální model (tedy model konstanty), dále v argumentu scope maximální model, který připadá v úvahu (zadá se pravá strana modelu, tedy v našem případě scope= ~ body.W + body.H + waist.C + hip.C + antb.C, a upřesnění směru direction='forward'. Metoda v prvním kroku zkouší přidat vždy jen jednu proměnnou a zjišťuje, jestli se přidáním konkrétní jedné proměnné model zlepší. Poté do modelu zahrne tu, která vede k největšímu zlepšení modelu. V dalším kroku pracuje s modelem, který ji obsahuje, a u něj opět zkouší přidávat ostatní proměnné. Opět vybere tu, jejíž vypuštění vede k největšímu zlepšení modelu, a takto pokračuje dále. Pokud metoda během kroku zjistí, že přidání jakékoli další proměnné nevede k zlepšení, proces končí.

```
model.for <- step(lm(neck.C ~ 1, data=neck),
                  scope= ~ body.W + body.H + waist.C + hip.C + antb.C, direction='forward')

## Start:  AIC=602.6
## neck.C ~ 1
##
##           Df Sum of Sq   RSS    AIC
## + antb.C   1     64034 22594 487.68
## + waist.C  1     62506 24123 493.37
## + body.W   1     58753 27875 505.95
## + body.H   1     33549 53080 561.99
## + hip.C    1     13611 73017 589.73
## <none>                 86628 602.60
##
## Step:  AIC=487.68
## neck.C ~ antb.C
##
##           Df Sum of Sq   RSS    AIC
## + waist.C  1    4317.8 18277 471.23
## + body.H   1    1873.3 20721 482.15
## + body.W   1    1688.5 20906 482.92
## <none>                 22594 487.68
## + hip.C    1     363.9 22231 488.27
##
## Step:  AIC=471.23
## neck.C ~ antb.C + waist.C
##
##          Df Sum of Sq   RSS    AIC
## + hip.C   1   3123.45 15153 456.92
## + body.H  1   2459.65 15817 460.65
## <none>                18277 471.23
## + body.W  1      0.08 18277 473.23
##
## Step:  AIC=456.92
## neck.C ~ antb.C + waist.C + hip.C
##
##          Df Sum of Sq   RSS    AIC
## + body.W  1    2327.3 12826 444.42
## + body.H  1    2060.8 13092 446.21
```

```
## <none>                15153 456.92
##
## Step:  AIC=444.42
## neck.C ~ antb.C + waist.C + hip.C + body.W
##
##          Df Sum of Sq   RSS    AIC
## + body.H  1    464.81 12361 443.21
## <none>                12826 444.42
##
## Step:  AIC=443.21
## neck.C ~ antb.C + waist.C + hip.C + body.W + body.H

model.for

##
## Call:
## lm(formula = neck.C ~ antb.C + waist.C + hip.C + body.W + body.H,
##     data = neck)
##
## Coefficients:
## (Intercept)       antb.C      waist.C        hip.C       body.W
##   165.63910      0.29120      0.18260     -0.18166      1.02594
##      body.H
##     0.04039
```

Vidíme, že pro náš model metoda skončila po šestém kroku. V prvním kroku přidala proměnnou antb.C, v dalším waist.C, v dalším hip.C, poté body.W, pak body.H a v šestém kroku již neměla k dispozici žadné další proměnné. Opět jsme tedy došli k modelu, který zahrnuje všechny proměnné.

V R je implementována i metoda both, která kombinuje metodu backward a forward. V jednotlivých krocích zkouší jak přidávání další proměnných (včetně těch, které byly v nějakém dřívějším kroku vypuštěny), tak vypouštění těch, které už v modelu máme. Prochází tedy veškeré možné kombinace vysvětlujících proměnných. Specifikovat ji můžeme tak, že zadáme maximální model a směr direction='both', v tom případě začne v prvním kroku proměnné vypouštět a v dalších krocích už bude zkoušet vypouštění i přidávání. Nebo jí můžeme spěcifikovat tím, že zadáme minimální model a v argumentu scope maximální model, a opět směr direction='both', v tomto případě bude v prvním kroku proměnné přidávat a v dalších krocích bude zkoušet přidávání i vypouštění.

```
model.both1 <- step(lm(neck.C ~ body.W + body.H + waist.C + hip.C + antb.C, data=neck),
                    direction='both')

## Start:  AIC=443.21
## neck.C ~ body.W + body.H + waist.C + hip.C + antb.C
##
##            Df Sum of Sq   RSS    AIC
## <none>                  12361 443.21
## - body.H    1    464.81 12826 444.42
## - antb.C    1    646.82 13008 445.64
## - body.W    1    731.29 13092 446.21
## - hip.C     1   3039.71 15401 460.33
## - waist.C   1   3140.65 15502 460.90

model.both1
```

```
##
## Call:
## lm(formula = neck.C ~ body.W + body.H + waist.C + hip.C + antb.C,
##     data = neck)
##
## Coefficients:
## (Intercept)       body.W       body.H      waist.C        hip.C
##   165.63910      1.02594      0.04039      0.18260     -0.18166
##      antb.C
##     0.29120
```

```
model.both2 <- step(lm(neck.C ~ 1, data=neck),
                    scope= ~ body.W + body.H + waist.C + hip.C + antb.C,
                    direction='both')
```

```
## Start:  AIC=602.6
## neck.C ~ 1
##
##           Df Sum of Sq   RSS    AIC
## + antb.C   1     64034 22594 487.68
## + waist.C  1     62506 24123 493.37
## + body.W   1     58753 27875 505.95
## + body.H   1     33549 53080 561.99
## + hip.C    1     13611 73017 589.73
## <none>                 86628 602.60
##
## Step:  AIC=487.68
## neck.C ~ antb.C
##
##           Df Sum of Sq   RSS    AIC
## + waist.C  1      4318 18277 471.23
## + body.H   1      1873 20721 482.15
## + body.W   1      1688 20906 482.92
## <none>                 22594 487.68
## + hip.C    1       364 22230 488.27
## - antb.C   1     64034 86628 602.60
##
## Step:  AIC=471.23
## neck.C ~ antb.C + waist.C
##
##           Df Sum of Sq   RSS    AIC
## + hip.C    1    3123.4 15153 456.92
## + body.H   1    2459.6 15817 460.65
## <none>                 18277 471.23
## + body.W   1       0.1 18277 473.23
## - waist.C  1    4317.8 22594 487.68
## - antb.C   1    5846.0 24123 493.37
##
## Step:  AIC=456.92
## neck.C ~ antb.C + waist.C + hip.C
##
##           Df Sum of Sq   RSS    AIC
## + body.W   1    2327.3 12826 444.42
## + body.H   1    2060.8 13092 446.21
```

```
## <none>                 15153 456.92
## - hip.C    1    3123.4 18277 471.23
## - antb.C   1    5140.1 20293 480.34
## - waist.C  1    7077.4 22231 488.27
##
## Step:  AIC=444.42
## neck.C ~ antb.C + waist.C + hip.C + body.W
##
##           Df Sum of Sq   RSS    AIC
## + body.H   1     464.8 12361 443.21
## <none>                 12826 444.42
## - antb.C   1     629.6 13456 446.59
## - body.W   1    2327.3 15153 456.92
## - waist.C  1    2679.0 15505 458.92
## - hip.C    1    5450.6 18277 473.23
##
## Step:  AIC=443.21
## neck.C ~ antb.C + waist.C + hip.C + body.W + body.H
##
##           Df Sum of Sq   RSS    AIC
## <none>                 12361 443.21
## - body.H   1    464.81 12826 444.42
## - antb.C   1    646.82 13008 445.64
## - body.W   1    731.29 13092 446.21
## - hip.C    1   3039.71 15401 460.33
## - waist.C  1   3140.65 15502 460.90
```

```
model.both2
```

```
##
## Call:
## lm(formula = neck.C ~ antb.C + waist.C + hip.C + body.W + body.H,
##     data = neck)
##
## Coefficients:
## (Intercept)       antb.C      waist.C        hip.C       body.W
##   165.63910      0.29120      0.18260     -0.18166      1.02594
##      body.H
##     0.04039
```