Regresja wieloliniowa i analiza kowariancji

W niniejszym zestawie zadań wykorzystujemy podzbiór danych (tylko jeden rok) pochodzących z badania the Child Health and Development Studies. Zawiera on informacje o 1236 urodzeniach chłopców, z których każdy żył co najmnij 28 dni. Przeprowadzono wywiady z matkami tych dzieci w trakcie których zebrano wiele medycznych i genetycznych danych, włączając w to informację czy matka jest palaczem tytoniu. Następnie po urodzeniu dzieci zostały zebrane dane o ich zdrowiu.

Zmiennne

zmienna	opis
`bwt`	waga dziecka po urodzeniu (uncje)
`gestation`	długość trwania ciąży liczona od ostatniej prawidłowej menstruacji (dni)
`parity`	indykator czy dziecko jest pierworodne (0 = nie, 1 = tak)
`age`	wiek matki w chwili poczęcia (lata)
`height`	wzrost matki (cale)
`weight`	waga matki (funty)
`smoke`	indykator czy matka jest palaczem (0 = nie, 1 = tak)

```
babies <- read.csv("C:/Users/Ewci/Desktop/wnioskowanie II/babies.csv")
```

Czyszczenie danych

Zadanie: zamieĹ„ odpowiednie zmienne na czynniki

```
factor(babies$smoke, levels=c(tak=1,nie=0), labels=c('tak','nie'))
```

```
##    [1] nie  nie  tak  nie  tak  nie  nie  nie  nie  tak  nie  tak  tak
##   [14] tak  nie  nie  tak  tak  nie  tak  nie  tak  nie  nie  nie  tak
##   [27] nie  tak  nie  tak  nie  nie  nie  nie  nie  nie  nie  tak  tak
##   [40] nie  nie  nie  tak  nie  tak  tak  nie  nie  nie  tak  nie  tak
##   [53] nie  nie  nie  nie  tak  nie  nie  nie  nie  nie  tak  tak  tak
##   [66] tak  nie  tak  tak  nie  nie  nie  tak  nie  tak  nie  tak  nie
##   [79] nie  tak  nie  nie  nie  tak  nie  tak  nie  nie  tak  tak  nie
##   [92] nie  nie  nie  tak  nie  tak  nie  nie  nie  tak  nie  nie  tak
##  [105] tak  nie  nie  nie  nie  nie  tak  nie  tak  nie  tak  nie  tak
##  [118] nie  nie  tak  nie  nie  nie  nie  nie  tak  nie  tak  nie  nie
##  [131] nie  tak  tak  nie  tak  nie  nie  nie  tak  nie  tak  nie  nie
##  [144] nie  tak  nie  nie  tak  tak  nie  nie  nie  nie  nie  tak  nie
##  [157] tak  tak  nie  nie  nie  nie  nie  nie  nie  tak  nie  tak  tak
##  [170] <NA> nie  tak  nie  nie  nie  tak  nie  tak  nie  nie  nie  tak
##  [183] nie  nie  nie  tak  nie  nie  nie  tak  tak  nie  tak  tak  tak
##  [196] nie  nie  tak  nie  nie  nie  nie  tak  nie  nie  nie  tak  tak
##  [209] nie  tak  tak  nie  tak  nie  nie  nie  nie  nie  <NA> tak  nie
##  [222] tak  nie  tak  nie  nie  nie  tak  tak  tak  tak  nie  tak  nie
##  [235] nie  tak  nie  nie  nie  nie  tak  tak  tak  tak  tak  nie  nie
##  [248] nie  nie  nie  tak  nie  nie  tak  <NA> <NA> nie  nie  nie  tak
##  [261] nie  tak  tak  nie  nie  tak  nie  nie  tak  nie  nie  nie  nie
##  [274] tak  tak  nie  tak  nie  nie  tak  tak  nie  tak  tak  tak  nie
##  [287] tak  nie  nie  nie  nie  tak  tak  tak  nie  tak  nie  tak  tak
##  [300] nie  nie  tak  nie  nie  nie  nie  tak  nie  tak  tak  tak  tak
##  [313] nie  tak  tak  tak  tak  tak  nie  nie  nie  nie  nie  nie  nie
##  [326] nie  tak  tak  nie  tak  nie  nie  tak  nie  nie  tak  tak  nie
##  [339] nie  tak  tak  tak  tak  nie  nie  nie  tak  tak  tak  nie  tak
##  [352] nie  tak  nie  tak  nie  nie  nie  tak  tak  nie  tak  nie  tak
##  [365] nie  nie  nie  nie  tak  nie  nie  nie  nie  tak  tak  tak  nie
##  [378] nie  tak  nie  nie  tak  nie  nie  nie  nie  tak  tak  tak  nie
##  [391] nie  tak  nie  nie  tak  tak  nie  tak  nie  nie  nie  nie  nie
##  [404] nie  nie  nie  tak  nie  nie  nie  nie  nie  nie  nie  nie  tak
##  [417] nie  tak  tak  nie  tak  nie  nie  tak  nie  nie  nie  tak  tak
##  [430] nie  nie  tak  <NA> tak  tak  nie  nie  tak  nie  nie  nie  nie
##  [443] tak  tak  nie  nie  nie  tak  tak  nie  nie  nie  nie  tak  nie
##  [456] tak  nie  tak  nie  tak  nie  nie  tak  tak  tak  tak  tak  nie
```

```
##  [456] tak  nie  tak  nie  tak  nie  nie  tak  tak  tak  tak  tak  nie
##  [469] tak  nie  nie  tak  nie  tak  tak  tak  tak  tak  nie  tak  tak
##  [482] nie  tak  tak  tak  nie  nie  tak  tak  nie  tak  tak  nie  nie
##  [495] nie  nie  nie  nie  tak  nie  tak  tak  tak  tak  tak  nie  nie
##  [508] nie  tak  tak  nie  nie  tak  nie  tak  tak  tak  tak  nie  tak
##  [521] nie  nie  tak  nie  nie  nie  tak  nie  tak  nie  nie  tak  tak
##  [534] nie  nie  tak  nie  nie  nie  nie  nie  nie  tak  nie  nie  tak
##  [547] nie  tak  tak  nie  nie  nie  nie  tak  nie  nie  nie  tak  nie
##  [560] nie  nie  tak  tak  nie  nie  nie  nie  tak  tak  tak  tak  nie
##  [573] tak  nie  tak  tak  nie  nie  nie  nie  nie  tak  tak  nie  nie
##  [586] tak  nie  nie  nie  tak  nie  tak  nie  tak  nie  nie  nie  tak
##  [599] nie  nie  <NA> tak  tak  tak  nie  nie  nie  tak  tak  tak  nie
##  [612] tak  tak  nie  nie  nie  nie  nie  tak  tak  nie  tak  nie  nie
##  [625] tak  nie  nie  nie  nie  nie  nie  tak  nie  nie  tak  tak  nie
##  [638] tak  tak  nie  tak  <NA> nie  tak  nie  nie  nie  tak  tak  nie
##  [651] nie  nie  nie  tak  nie  tak  tak  nie  nie  tak  tak  nie  tak
##  [664] nie  tak  <NA> tak  tak  tak  nie  tak  <NA> tak  nie  nie  tak
##  [677] nie  tak  nie  nie  nie  nie  tak  nie  nie  nie  nie  nie  tak
##  [690] nie  tak  nie  nie  tak  tak  nie  tak  tak  tak  nie  tak  nie
##  [703] nie  nie  nie  nie  tak  nie  nie  nie  nie  tak  tak  nie  tak
##  [716] tak  tak  tak  nie  nie  nie  nie  nie  nie  nie  nie  tak  nie
##  [729] nie  nie  nie  nie  nie  nie  tak  tak  nie  tak  nie  nie  nie
##  [742] tak  nie  nie  tak  nie  nie  nie  nie  nie  tak  nie  nie  tak
##  [755] nie  tak  tak  nie  nie  tak  tak  tak  tak  nie  tak  nie  nie
##  [768] nie  nie  nie  tak  nie  nie  nie  tak  nie  nie  nie  nie  nie
##  [781] tak  nie  nie  nie  tak  tak  nie  nie  nie  nie  tak  nie  tak
##  [794] tak  nie  nie  nie  nie  tak  tak  nie  tak  tak  nie  tak  nie
##  [807] tak  nie  tak  nie  nie  tak  tak  nie  nie  nie  tak  nie  tak
##  [820] nie  tak  nie  nie  nie  nie  nie  nie  nie  tak  nie  tak  tak
##  [833] tak  tak  nie  nie  nie  nie  nie  nie  nie  nie  nie  nie  nie
##  [846] tak  nie  tak  tak  nie  nie  tak  tak  nie  tak  nie  nie  nie
##  [859] nie  tak  tak  nie  nie  nie  nie  nie  nie  tak  tak  nie  nie
##  [872] nie  nie  nie  nie  nie  nie  tak  tak  nie  nie  nie  tak  tak
##  [885] nie  tak  nie  nie  nie  nie  nie  nie  nie  tak  nie  nie  tak
##  [898] nie  nie  nie  nie  nie  tak  tak  nie  tak  nie  tak  tak  nie
##  [911] nie  nie  tak  tak  nie  nie  nie  nie  tak  nie  <NA> nie  nie
##  [924] tak  nie  tak  nie  tak  tak  nie  tak  tak  nie  tak  tak  tak
##  [937] nie  tak  nie  tak  tak  nie  nie  nie  nie  nie  tak  nie  nie
##  [950] tak  nie  nie  tak  tak  nie  tak  nie  tak  nie  nie  nie  tak
##  [963] nie  nie  nie  tak  tak  tak  nie  tak  nie  nie  nie  nie  nie
##  [976] nie  tak  nie  nie  tak  nie  tak  nie  nie  tak  tak  nie  tak
##  [989] tak  nie  tak  nie  nie  nie  tak  nie  nie  nie  nie  tak  nie
## [1002] nie  tak  nie  tak  tak  tak  tak  nie  tak  tak  nie  tak  nie
## [1015] nie  tak  nie  tak  nie  nie  tak  nie  tak  nie  nie  nie  nie
## [1028] tak  nie  nie  nie  nie  nie  nie  nie  nie  nie  nie  nie  nie
## [1041] tak  nie  nie  tak  nie  nie  tak  nie  nie  nie  tak  tak  nie
## [1054] nie  nie  tak  nie  tak  tak  tak  nie  tak  tak  nie  tak  tak
## [1067] nie  nie  nie  tak  nie  tak  tak  nie  nie  nie  nie  nie  nie
## [1080] nie  nie  nie  nie  nie  nie  nie  tak  nie  tak  tak  tak  nie
## [1093] nie  nie  nie  nie  nie  nie  nie  nie  nie  tak  tak  nie  nie
## [1106] tak  tak  tak  nie  tak  nie  nie  nie  nie  nie  nie  nie  nie
## [1119] tak  nie  nie  nie  nie  tak  nie  nie  nie  tak  nie  nie  tak
## [1132] nie  tak  tak  nie  tak  tak  nie  nie  nie  nie  tak  nie  tak
## [1145] nie  tak  nie  tak  tak  tak  tak  nie  tak  nie  tak  nie  nie
## [1158] nie  nie  tak  nie  tak  tak  nie  tak  nie  nie  nie  tak  tak
## [1171] tak  nie  nie  nie  tak  tak  tak  nie  nie  tak  nie  nie  tak
## [1184] nie  nie  nie  tak  tak  nie  tak  nie  tak  nie  nie  tak  tak
## [1197] tak  tak  tak  nie  tak  nie  nie  tak  nie  nie  tak  tak  tak
## [1210] nie  tak  tak  nie  nie  nie  tak  nie  nie  tak  nie  nie  tak
## [1223] tak  tak  tak  tak  tak  tak  nie  nie  nie  nie  nie  tak  nie
## [1236] nie
## Levels: tak nie
```

```
factor(babies$parity,levels=c(tak=1,nie=0),labels=c('tak','nie'))
```

```
##    [1] nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie
##   [18] nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie
##   [35] nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie
##   [52] nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie
##   [69] nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie
##   [86] nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie
##  [103] nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie
```

```
##  [120] nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie
##  [137] nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie
##  [154] nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie
##  [171] nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie
##  [188] nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie
##  [205] nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie
##  [222] nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie
##  [239] nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie
##  [256] nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie tak
##  [273] nie tak nie nie tak nie nie tak nie tak nie tak tak tak tak nie tak
##  [290] nie tak nie tak tak tak nie nie nie tak nie tak nie nie tak nie tak
##  [307] nie tak nie nie nie nie nie nie nie nie nie nie nie nie tak nie tak
##  [324] nie nie tak nie nie tak nie nie nie tak tak tak nie tak nie nie nie
##  [341] tak nie nie nie nie nie nie nie nie tak nie nie nie nie tak tak nie nie
##  [358] nie nie nie tak nie nie nie nie nie tak nie nie nie nie tak nie nie nie
##  [375] nie nie tak nie nie tak nie tak nie nie nie nie tak nie nie tak nie
##  [392] tak tak nie tak nie tak tak nie tak nie nie nie tak tak tak nie nie
##  [409] tak nie nie nie tak nie tak nie nie nie nie tak nie tak tak nie nie
##  [426] nie nie tak nie tak nie tak nie nie nie nie nie nie nie nie nie tak
##  [443] nie nie nie nie tak tak nie tak tak tak nie nie nie nie tak nie nie
##  [460] nie tak tak nie tak nie tak nie tak tak tak nie nie tak nie tak tak
##  [477] nie tak nie tak nie nie nie nie nie nie nie tak nie tak nie nie nie tak
##  [494] tak nie tak nie tak nie nie nie nie nie tak tak nie tak nie tak nie
##  [511] nie tak nie nie tak nie nie nie nie nie nie nie nie nie nie nie tak
##  [528] nie tak nie nie nie nie nie nie nie tak nie tak tak tak nie nie nie
##  [545] tak tak nie nie tak tak nie nie nie nie tak tak nie nie tak tak nie
##  [562] nie tak nie nie nie tak nie nie nie nie nie tak tak nie nie nie nie
##  [579] nie nie nie tak nie tak nie tak nie tak nie nie nie nie nie nie tak
##  [596] nie tak tak tak tak nie tak tak nie nie nie tak tak tak nie tak nie
##  [613] nie tak tak nie nie nie nie tak nie nie nie nie nie nie nie tak nie
##  [630] tak tak tak tak nie tak nie nie tak nie nie nie nie nie tak nie nie
##  [647] nie nie nie tak nie nie tak nie nie nie nie nie nie nie nie nie nie
##  [664] nie tak nie nie tak nie tak nie nie nie nie nie nie nie nie nie nie
##  [681] nie nie nie tak nie tak nie nie tak nie nie tak nie nie nie nie nie
##  [698] nie nie tak tak nie tak nie nie tak nie tak tak nie nie tak nie nie
##  [715] nie nie tak tak nie nie tak tak nie nie nie nie nie nie nie nie tak
##  [732] nie nie nie nie nie tak nie tak nie tak nie tak tak nie tak tak nie
##  [749] nie nie nie nie nie nie nie nie tak nie nie tak nie nie nie nie nie
##  [766] tak nie nie nie tak nie nie tak nie nie nie nie nie tak nie nie tak
##  [783] nie tak nie nie nie nie nie tak nie nie tak tak nie nie tak nie tak
##  [800] nie nie nie tak nie nie nie tak tak nie nie nie nie nie nie nie nie
##  [817] nie nie tak nie nie nie nie nie tak tak nie tak tak nie nie tak nie
##  [834] nie nie nie nie tak tak tak nie tak nie tak nie tak nie nie nie nie
##  [851] tak nie nie nie nie nie nie nie nie nie tak tak nie tak nie nie tak
##  [868] nie nie nie tak nie tak nie nie tak tak tak nie tak tak tak tak tak
##  [885] nie nie tak nie tak nie tak nie tak nie nie nie tak nie tak nie tak
##  [902] nie tak nie nie tak nie nie nie tak tak tak nie tak tak nie nie nie
##  [919] tak tak nie nie nie nie nie tak nie nie tak tak tak nie tak tak nie
##  [936] nie nie tak nie tak nie tak nie nie tak nie nie tak nie tak nie nie
##  [953] nie nie nie tak nie nie nie tak nie nie nie nie nie nie nie nie tak
##  [970] nie tak nie tak tak tak nie tak nie nie tak nie tak nie nie nie tak
##  [987] tak nie nie nie nie tak tak tak nie nie nie tak nie tak nie nie nie
## [1004] nie tak tak tak nie nie nie tak nie nie nie nie tak nie nie nie nie
## [1021] nie nie nie nie nie nie nie tak nie nie nie tak nie nie nie nie nie
## [1038] nie tak nie nie nie tak nie nie tak nie nie tak nie tak tak tak tak
## [1055] nie nie nie nie nie tak nie nie nie nie tak nie tak nie nie tak tak
## [1072] nie nie tak nie nie nie nie tak tak tak nie nie tak nie nie nie tak
## [1089] tak tak nie tak nie tak nie nie nie nie tak nie nie nie tak nie tak
## [1106] nie nie tak nie nie nie nie nie nie nie tak nie nie nie nie nie nie nie
## [1123] nie nie nie tak tak nie nie nie nie nie nie tak nie nie nie tak tak
## [1140] tak nie nie tak nie nie nie nie nie nie tak nie nie nie nie tak nie nie
## [1157] nie nie nie nie nie nie nie nie tak tak nie nie nie nie nie tak nie
## [1174] nie tak nie nie nie tak nie nie nie nie nie nie tak nie nie nie tak tak
## [1191] nie nie tak nie nie nie nie nie tak nie nie tak nie nie tak nie nie nie
## [1208] tak nie tak nie nie tak nie nie nie nie nie nie nie nie tak tak tak tak
## [1225] nie tak tak nie nie nie nie tak nie nie tak nie
## Levels: tak nie
```

W naszym zbiorze danych występuje trochę niepoprawnych wartości (np. mało prawdopodobne jest by ciąża trwała 999 dni). W związku z tym występuje potrzeba odpowiedniego wyczyszczenia danych, i zastąpienia błędnych danych wartością NA. W szczególności warto

przyjrzeć się zmiennym `gestation`, `age`, `weight`, `height`

Zadanie: zastąp nierealne wartości zmiennych wartością NA

```
babies[babies$weight==999,"weight"]<-NA
babies[babies$gestation==999,"gestation"]<-NA
babies[babies$smoke==9,"smoke"]<-NA
```

Żeby łatwiej było interpretować wyniki, dobrze jest zamienić występujące wartości na system jednostek SI.

Zadanie: zamień wartości, tak aby odpowiednie zmienne były reprezentowane w systemie jednostek SI

```
babies$weight=0.45*babies$weight
babies$bwt=0.028*babies$bwt
```

Podstawowe statystyki

Zadanie: wyświetl i skomentuj podstawowe statystki dla zmiennych w naszym zbiorze danych*

```
model=lm(bwt ~ height+weight, babies)
summary(babies)
```

```
##       bwt          gestation         parity           age       
##  Min.   :1.540   Min.   :148.0   Min.   :0.0000   Min.   :15.00  
##  1st Qu.:3.045   1st Qu.:272.0   1st Qu.:0.0000   1st Qu.:23.00  
##  Median :3.360   Median :280.0   Median :0.0000   Median :26.00  
##  Mean   :3.348   Mean   :279.3   Mean   :0.2549   Mean   :27.37  
##  3rd Qu.:3.668   3rd Qu.:288.0   3rd Qu.:1.0000   3rd Qu.:31.00  
##  Max.   :4.928   Max.   :353.0   Max.   :1.0000   Max.   :99.00  
##                  NA's   :13                                      
##      height          weight           smoke       
##  Min.   :53.00   Min.   : 39.15   Min.   :0.0000  
##  1st Qu.:62.00   1st Qu.: 51.64   1st Qu.:0.0000  
##  Median :64.00   Median : 56.25   Median :0.0000  
##  Mean   :64.67   Mean   : 57.88   Mean   :0.3948  
##  3rd Qu.:66.00   3rd Qu.: 62.55   3rd Qu.:1.0000  
##  Max.   :99.00   Max.   :112.50   Max.   :1.0000  
##                  NA's   :36       NA's   :10
```

Regresja wieloliniowa

Zazwyczaj staramy się wyjaśnić wartości zmiennej zależnej za pomocą więcej niż jednego predyktora. Mówimy wtedy, że mamy do czynienia z regresją wieloliniową (ang.multiple linear regression). Dla przykładu rozważmy model, który wyjaśnia zmienną `bwt` za pomocą zmiennych `height` i `weight`

```
summary(lm(bwt ~ height+weight, babies))
```

```
##
## Call:
## lm(formula = bwt ~ height + weight, data = babies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85522 -0.30022  0.01399  0.31479  1.57264
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.471050   0.304876   4.825 1.58e-06 ***
## height      0.024182   0.005069   4.771 2.06e-06 ***
## weight      0.005592   0.001649   3.391 0.000718 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5036 on 1197 degrees of freedom
##   (36 observations deleted due to missingness)
## Multiple R-squared:  0.04195,    Adjusted R-squared:  0.04034
## F-statistic:  26.2 on 2 and 1197 DF,  p-value: 7.278e-12
```

Powyższy model przewiduje, że nasze dane opisywane są zależnością

$$\text{bwt} = a \cdot \text{height} + b \cdot \text{weight} + c$$

Podobnie jak w przypadku regresji prostej współczynniki estymowane są za pomocą metody najmniejszych kwadratów.

Zadanie: Jakie są estymacje współczynników a, b i c w powyższym modelu? Jak można zinterpretować powyższe współczynniki?

```
model$coefficients
```

```
## (Intercept)      height      weight
##  1.47104978  0.02418245  0.00559229
```

Analiza wariancji

Możemy starać się przewidywać wartość zmiennej `bwt` za pomocą zmiennej `smoke`, która wskazuje na to czy matka jest nałogowym palaczem, czy nie. Zwróćmy jednak uwagę na istotną różnicę w stosunku do regresji prostej. Mianowicie, zmienna `smoke` nie jest liczbowa. W takiej sytuacji, gdy zmienną numeryczną przewidujemy za pomocą zmiennej kategorycznej, mówimy, że mamy do czynienia analizą wariancji (ANOVA).

Zauważmy, że aby dopasować model ANOVA, nie musimy tworzyć jednakże nowej teorii matematycznej. Wystarczy zamienić czynnik o dwóch poziomach na zmienną liczbową wprowadzając tzw. indykator. Zatem niech e_{smoker} oznacza zmiennną, która przyjmuje wartość 1, gdy matka jest palaczem i wartość 0 w przeciwnym przypadku. Mając do dyspozycji zmienną liczbową, możemy zrobić regresję ze względu na indykator e_{smoker}, tzn. dopasować model liniowy

$$\text{bwt} = a \cdot e_{\text{smoker}} + b$$

Często zapisuje się go w następującej postaci

$$\text{bwt} = a \cdot \text{smoker} + b$$

co nie ma większego sensu matematycznego (bo `smoker` nie jest zmienną liczbową, a jedynie poziomem zmiennej `smoke`). Natomiast, rozumiemy to jako pewien skrót myślowy, wstawiając w miejsce `smoker` odpowiedni indykator.

W związku z powyższym interpretacja współczynników jest następująca. Jeżeli matka jest palaczem, średnia waga dziecka wynosi $a + b$ (wstawiamy za indykator wartość 1). Natomiast w przeciwnym przypadku średnia waga noworodka wynosi a.

Model ANOVA możemy dopasować do naszych danych w sposób analogiczny do zwykłej regresji

```
summary(lm(bwt~smoke, babies))
```

```
##
## Call:
## lm(formula = bwt ~ smoke, data = babies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.90532 -0.30932  0.02493  0.30668  1.48268
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.44532    0.01817 189.597   <2e-16 ***
## smoke       -0.25025    0.02892  -8.653   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.495 on 1224 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.05764,    Adjusted R-squared:  0.05687
## F-statistic: 74.87 on 1 and 1224 DF,  p-value: < 2.2e-16
```

Zadanie: zinterpretuj powyższe wyniki

Jeżeli kobieta pali to waga dziecka będzie mniejsza, niz waga dziecka urodzonego przez kobietę, która nie pali. . ***Smokesmoker to wspÄłĹ‚czynnik, ktÄłry objaĹ›nia o ile mniejszÄ… wagÄ™ bÄ™dzie miaĹ‚o dziecko urodzone przez matkÄ™ palÄ…cÄ… od wagi dziecka,ktÄłrego matkajest osoba niepalÄ…cÄ…. Intercept to wspÄłĹ‚czynnik, ktory oznacza wagÄ™ dziecka urodzonego przez matkÄ™ niepalÄ…cÄ….***

Analiza kowariancji

Zmienną numeryczną możemy wyjaśniać za pomocą innej zmiennej numerycznej i zmiennej kategorycznej jednocześnie. Tak jak w poniższym przykładzie

```
model.height.smoke <- lm(bwt~height+smoke, babies)
```

Jest to sytuacja która jest połączeniem regresji wieloliniowej i analizy wariancji omówionej pokrótce w poprzedniej sekcji. W takim przypadku przyjęło się używać nazwy *analiza kowariancji* (ANCOVA).

Warto podkreślić, że używamy tutaj różnych nazw (analiza regresji/wariancji/kowariancji) dla w zasadzie tego samego matematycznego mechanizmu: tzn. dopasowania funkcji liniowej do naszego zbioru danych. Dlatego wiele osób nie rozróżnia pomiędzy tymi różnymi rodzajami analiz, mówiąc po prostu o modelach liniowych. Jednak z drugiej strony, różne rodzaje analiz używane są w różnch kontekstach badawczych, dlatego przyjęły się różne nazwy.

W następnej sekcji bliżej przyjżymy się powyższemu modelowi ANCOVA.

Jakie są predyktory wagi dziecka?

Jednym z dobrych predyktorów wagi dziecka jest wzrost matki. Rzecz jasna wyższe matki typowo rodzą cięższe dzieci. Zależność możemy zaobserwować na poniższym wykresie

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
library(dplyr)
```

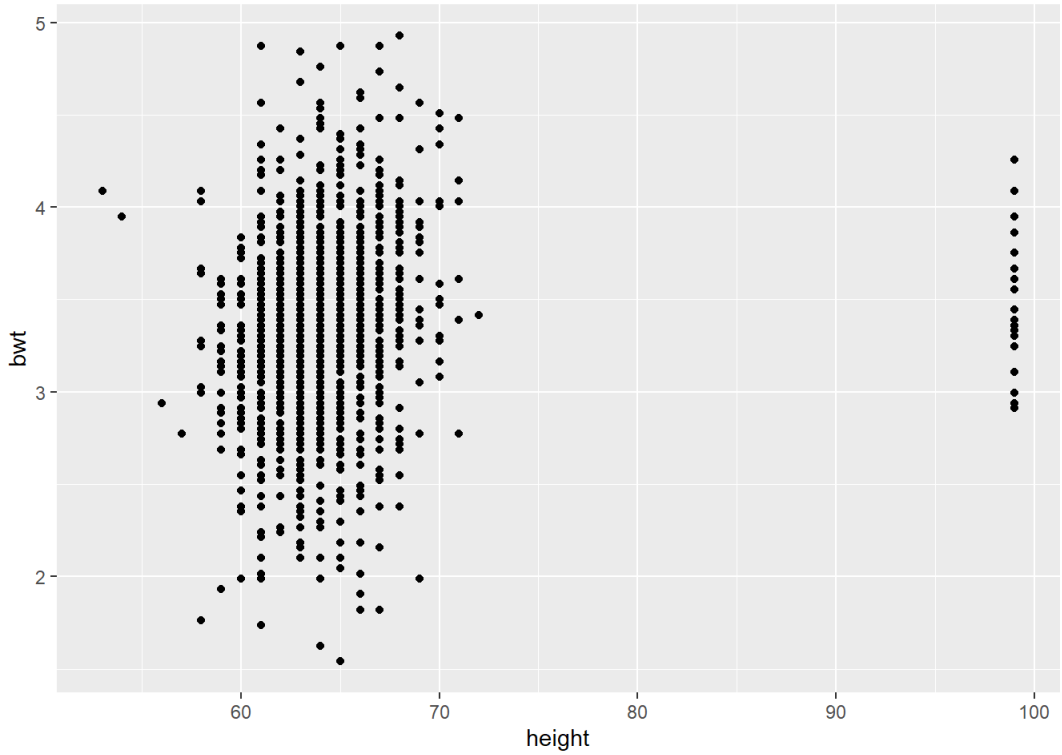
```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
ggplot(babies, aes(x = height, y = bwt)) + geom_point()
```




```
babies %>% filter(!is.na(height) & !is.na(bwt)) %>% summarise(correlation = cor(height, bwt))
```

correlation <dbl>
0.1255413

1 row

Widzimy, że współczynnik korelacji jest istotny (około 0.2), jednakże nie jest bardzo duży, co sugeruje, że na wagę dziecka wpływają również inne czynniki.

Zadanie: za pomocą funkcji `ggpairs` z pakietu `GGally` sprawdź jak wyglądają korelacje `bwt` z innymi zmiennymi numerycznymi

---

```
library("GGally")
```

```
## Warning: package 'GGally' was built under R version 3.5.3
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
ggpairs(babies[1:6])
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 13 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 36 rows containing missing values
```

```
## Warning: Removed 13 rows containing missing values (geom_point).
```

```
## Warning: Removed 13 rows containing non-finite values (stat_density).
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 13 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 13 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 13 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 48 rows containing missing values
```

```
## Warning: Removed 13 rows containing missing values (geom_point).
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 36 rows containing missing values
```

```
## Warning: Removed 13 rows containing missing values (geom_point).
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 36 rows containing missing values
```

```
## Warning: Removed 13 rows containing missing values (geom_point).
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 36 rows containing missing values
```

```
## Warning: Removed 36 rows containing missing values (geom_point).
```

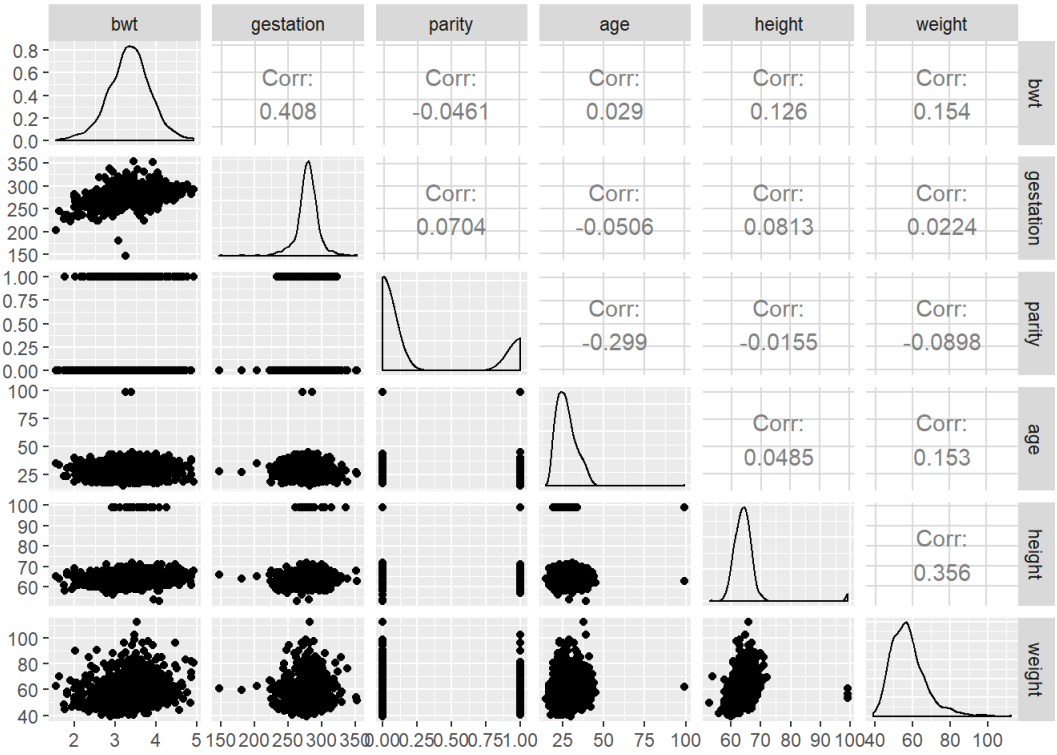
```
## Warning: Removed 48 rows containing missing values (geom_point).
```

```
## Warning: Removed 36 rows containing missing values (geom_point).

## Warning: Removed 36 rows containing missing values (geom_point).

## Warning: Removed 36 rows containing missing values (geom_point).
```

```
## Warning: Removed 36 rows containing non-finite values (stat_density).
```



Największa korelację między wagą a inną zmienną numeryczną widzimy dla zmiennej gestation. Kolejnna zmienna wysoko skorelowana ze zmienną waga jest to zmienna wagi kobiety.

***

Według amerykańskich służb zdrowia publicznego (United States Public Health Service) to czy matka jest nałogowym palaczem jest czynnikiem wpływającym na wagę dziecka bardziej niż wzrost matki. W tej części ćwiczenia zbadamy to stwierdzenie.

Zadanie: zbadaj powyższe stwierdzenie dopasowując do zmiennej `bwt` dwa modele liniowe. Pierwszy niech wyjaśnia wagę dziecka za pomocą wzrostu matki, a drugi za pomocą zmiennej `smoke`. Porównaj estymacje wariancji dla obu tych modeli. Ponadto na podstawie współczynnika determinacji r^2 oblicz procent wariancji wyjaśnianej przez poszczególne predyktory

---

```
summary(lm(bwt~height, babies))
```

```
##
## Call:
## lm(formula = bwt ~ height, data = babies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81217 -0.31236  0.01638  0.31946  1.56856
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.560175   0.177849  14.395  < 2e-16 ***
## height      0.012185   0.002741   4.445 9.57e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5068 on 1234 degrees of freedom
## Multiple R-squared:  0.01576,    Adjusted R-squared:  0.01496
## F-statistic: 19.76 on 1 and 1234 DF,  p-value: 9.569e-06
```

```
summary(lm(bwt~smoke, babies))
```

```
##
## Call:
## lm(formula = bwt ~ smoke, data = babies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.90532 -0.30932  0.02493  0.30668  1.48268
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.44532    0.01817 189.597   <2e-16 ***
## smoke       -0.25025    0.02892  -8.653   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.495 on 1224 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.05764,    Adjusted R-squared:  0.05687
## F-statistic: 74.87 on 1 and 1224 DF,  p-value: < 2.2e-16
```

```
lm(bwt~height, babies)
```

```
##
## Call:
## lm(formula = bwt ~ height, data = babies)
##
## Coefficients:
## (Intercept)       height
##     2.56018      0.01218
```

```
lm(bwt~smoke, babies)
```

```
##
## Call:
## lm(formula = bwt ~ smoke, data = babies)
##
## Coefficients:
## (Intercept)        smoke
##      3.4453      -0.2503
```

Estymacja wariancji to 508.5, 501.2. Dopasowaliśmy już model, który używa dwóch zmiennych na raz do “wyjaśniania” zmienności wagi ciała noworodków. Przeanalizujmy jego wyniki

```
summary(model.height.smoke)
```

```
##
## Call:
## lm(formula = bwt ~ height + smoke, data = babies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.91291 -0.30419  0.01133  0.31352  1.47338
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.570693   0.172149  14.933  < 2e-16 ***
## height       0.013573   0.002657   5.109 3.76e-07 ***
## smoke       -0.258238   0.028672  -9.007  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.49 on 1223 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.07733,    Adjusted R-squared:  0.07582
## F-statistic: 51.25 on 2 and 1223 DF,  p-value: < 2.2e-16
```

Jak można było oczekiwać model zawierający dwie zmienne niezależne jest lepiej dopasowany od modeli z pojedynczymi zmiennymi.

Zadanie: jaki procent wariancji “wyjaśnia” model z dwoma zmiennymi. Jaki jest związek z wariancją “wyjaśnianą” przez modele z pojedynczymi zmiennymi

---

```
mod=lm(bwt~height, babies)$coef
mod2=lm(bwt~smoke, babies)$coef
```

Model z dwoma zmiennymi wyjaĹ›nia 10.24% wariancji ***

Dopasujmy teraz model, który uwzględnia również trzecią zmienną, mianowicie wagę matki

```
model.height.smoke.weight <- lm(bwt~height+weight+smoke, babies)
summary(model.height.smoke.weight)
```

```
##
## Call:
## lm(formula = bwt ~ height + weight + smoke, data = babies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94855 -0.27973  0.00623  0.31003  1.46927
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.459820   0.295985   4.932 9.29e-07 ***
## height       0.026903   0.004927   5.460 5.79e-08 ***
## weight       0.004445   0.001612   2.758  0.00591 **
## smoke       -0.251677   0.029075  -8.656  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4876 on 1186 degrees of freedom
##   (46 observations deleted due to missingness)
## Multiple R-squared:  0.09956,    Adjusted R-squared:  0.09728
## F-statistic: 43.71 on 3 and 1186 DF,  p-value: < 2.2e-16
```

Interakcje pozwalają lepiej modelować zależności pomiędzy predyktorami. Wyjaśnimy to na przykładzie. Wiemy, że masa ciała noworodka wzrasta wraz ze wzrostem matki, potrafimy nawet wyliczyć współczynnik tego wzrostu. Wiemy, też że masa ciała noworodków zależy od tego czy matka jest palaczem, czy nie.

Może się jednak okazać, że współczynnik określający zmianę przeciętnej masy ciała noworodka w zależności od wzrostu matki zależy od tego czy matka pali (być może wzrost matki jest słabszym predyktorem `bwt` dla palących matek?). Możemy do tej kwestii się odnieść, dopasowując poniższy model z interakcjami

```
model.height.smoke.int <- lm(bwt~height*smoke, babies)
summary(model.height.smoke.int)
```

```
##
## Call:
## lm(formula = bwt ~ height * smoke, data = babies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9128 -0.3045  0.0110  0.3137  1.4730
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.5782298  0.2554067  10.095  < 2e-16 ***
## height        0.0134556  0.0039536   3.403 0.000687 ***
## smoke        -0.2720479  0.3467536  -0.785 0.432865
## height:smoke  0.0002134  0.0053405   0.040 0.968130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4902 on 1222 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.07733,    Adjusted R-squared:  0.07507
## F-statistic: 34.14 on 3 and 1222 DF,  p-value: < 2.2e-16
```

Faktycznie, możemy zaobserwować, że współczynnik przy zmiennej `height` będzie inny dla matek niepalących (17.2 grama) a inny dla matek palących (około 15.5 gramów). Paradoksalnie, palenie papierosów jest teraz powiązane z wagą większą o 12 gramów! Ale dlaczego tak jest wyjaśnia następujące zadanie.

Zadanie: według powyższego modelu oblicz średnią wagę noworodka dla matki o przeciętnym wzroście. O ile różni się ta waga dla matek palących i niepalących?

```
y1 = 17.177*median(babies$height[babies$smoke==0], na.rm=TRUE) + 694.867
y2 = (17.177-12.424)*median(babies$height[babies$smoke==1], na.rm=TRUE) + 694.867
abs(y1-y2)
```

```
## [1] 795.136
```

Model o największej zdolności predykcji

Jak widać, możemy dopasować wiele modeli wyjaśniających zmienność masy ciała noworodków. W zależności od zastosowań zasadne jest używanie różnych kryteriów wyboru modelu.

Generalnie zawsze staramy się, by model był jak najprostszy, gdyż nie zależy nam by dopasował się "zbyt dobrze" do naszych danych. Takie zbyt dobre dopasowanie może mieć następujące negatywne konsekwencje: jeżeli zależy nam na badaniu statystycznych zależności między zmiennymi (np. w celach naukowych) zbyt dobre dopasowanie może uznać za znaczące zależności, których w naszej populacji faktycznie nie ma. Natomiast, jeżeli zależy nam na zdolności predykcyjnej naszego modelu, może okazać się, że model zbyt dobrze dopasowany będzie źle działał na nowym zbiorze danych.

Jednym z typowych kryteriów używanych do wyboru modelu jest znany nam już zmodyfikowany współczynnik determinacji. Zawiera on w sobie informację o liczbie parametrów i w związku z tym nie zawsze rośnie przy bardziej skomplikowanych modelach. Inną metodą jest wybór odpowiedniego modelu w oparciu o p-wartości. Niestety, jak widzieliśmy może to powodować problemy w przypadku zmiennych, które są ze sobą skorelowane.

Generalnie uważa się, że od powyższych metod lepsze są tzw. *kryteria informacyjne*. Estymują one informację traconą w skutek używania naszego modelu, a zatem oczekujemy dla nich jak najmniejszej wartości. Typowo stosowane są: Bayesowskie kryterium informacyjne Schwarza (BIC) oraz kryterium informacyjne Akaikego (AIC). Formuły określające oba kryteria są bardzo podobne

$$BIC = \ln(n)p - 2\ln(\hat{L})$$

$$AIC = 2p - 2\ln(\hat{L})$$

gdzie p oznacza liczbę parametrów modelu, a $\hat{L}$ jest estymowaną wartością funkcji wiarygodności. BIC jest typowo bardziej konserwatywne niż AIC, co oznacza, że lepiej nadaje się do wykorzystania w sytuacji, gdy zależy nam przede wszystkim na istotnych statystycznie predyktorach. Natomiast AIC będzie sprawdzać się lepiej w modelach predykcyjnych.

W naszych poszukiwaniach najlepszego modelu używać będziemy funkcji `step`, która bada kryteria informacyjne (domyślnie AIC), eliminując poszczególne predyktory krok po kroku.

Generalnie, przy wyborze najlepszego modelu, staramy zaczynać nasze rozważania od modelu, który zawiera jak najwięcej predyktorów (część z nich zostanie później wyeliminowana przez funkcję `step`). Jednakże, jak się okazuje, trzeba przy wyborze takiego pełnego modelu również wykazać się pewnym rozsądkiem i nie brać wszystkich możliwych predyktorów.

Zaprezentujemy to zjawisko na poniższym przykładzie. Rozważymy pełny model, z wszystkimi możliwymi interakcjami.

```
model.full <- lm(bwt ~ height*weight*age*gestation*smoke*parity, na.omit(babies))
summary(model.full)
```

```
##
## Call:
## lm(formula = bwt ~ height * weight * age * gestation * smoke *
##     parity, data = na.omit(babies))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57003 -0.27992  0.00092  0.26172  1.37567
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             1.261e+01  3.348e+02   0.038
## height                                 -1.055e-01  5.293e+00  -0.020
## weight                                 -4.255e-01  5.738e+00  -0.074
## age                                    -1.601e+00  1.203e+01  -0.133
## gestation                              -1.022e-01  1.207e+00  -0.085
## smoke                                  -8.156e+02  4.831e+02  -1.688
## parity                                 -1.900e+03  1.236e+03  -1.537
## height:weight                           7.196e-03  9.037e-02   0.080
## height:age                              1.985e-02  1.901e-01   0.104
## weight:age                              3.392e-02  2.039e-01   0.166
## height:gestation                        1.467e-03  1.907e-02   0.077
## weight:gestation                        2.584e-03  2.065e-02   0.125
## age:gestation                           7.845e-03  4.329e-02   0.181
## height:smoke                            1.251e+01  7.570e+00   1.653
## weight:smoke                            1.360e+01  8.303e+00   1.638
## age:smoke                               3.076e+01  1.747e+01   1.760
## gestation:smoke                         3.018e+00  1.738e+00   1.736
## height:parity                           2.956e+01  1.915e+01   1.543
## weight:parity                           2.823e+01  2.151e+01   1.313
## age:parity                              7.083e+01  4.940e+01   1.434
## gestation:parity                        6.954e+00  4.310e+00   1.613
## smoke:parity                            2.763e+03  1.650e+03   1.674
## height:weight:age                      -5.102e-04  3.211e-03  -0.159
## height:weight:gestation                -4.234e-05  3.250e-04  -0.130
## height:age:gestation                   -1.045e-04  6.838e-04  -0.153
## weight:age:gestation                   -1.572e-04  7.334e-04  -0.214
## height:weight:smoke                    -2.100e-01  1.298e-01  -1.619
## height:age:smoke                       -4.734e-01  2.741e-01  -1.727
## weight:age:smoke                       -5.086e-01  2.998e-01  -1.696
## height:gestation:smoke                 -4.635e-02  2.723e-02  -1.702
## weight:gestation:smoke                 -5.009e-02  2.976e-02  -1.683
## age:gestation:smoke                    -1.139e-01  6.299e-02  -1.808
## height:weight:parity                   -4.397e-01  3.326e-01  -1.322
## height:age:parity                      -1.101e+00  7.637e-01  -1.442
## weight:age:parity                      -1.018e+00  8.545e-01  -1.191
## height:gestation:parity                -1.082e-01  6.676e-02  -1.620
## weight:gestation:parity                -1.044e-01  7.506e-02  -1.391
## age:gestation:parity                   -2.587e-01  1.720e-01  -1.504
## height:smoke:parity                    -4.357e+01  2.566e+01  -1.698
## weight:smoke:parity                    -4.511e+01  3.045e+01  -1.481
## age:smoke:parity                       -1.124e+02  6.600e+01  -1.703
## gestation:smoke:parity                 -9.965e+00  5.794e+00  -1.720
## height:weight:age:gestation             2.390e-06  1.154e-05   0.207
## height:weight:age:smoke                 7.855e-03  4.690e-03   1.675
## height:weight:gestation:smoke           7.736e-04  4.649e-04   1.664
## height:age:gestation:smoke              1.754e-03  9.882e-04   1.775
## weight:age:gestation:smoke              1.876e-03  1.078e-03   1.741
## height:weight:age:parity                1.583e-02  1.317e-02   1.202
## height:weight:gestation:parity          1.625e-03  1.159e-03   1.402
## height:age:gestation:parity             4.019e-03  2.657e-03   1.513
## weight:age:gestation:parity             3.762e-03  2.978e-03   1.263
## height:weight:smoke:parity              7.114e-01  4.701e-01   1.513
```

```
## height:age:smoke:parity                   1.777e+00  1.026e+00   1.732
## weight:age:smoke:parity                   1.838e+00  1.228e+00   1.496
## height:gestation:smoke:parity             1.572e-01  9.005e-02   1.746
## weight:gestation:smoke:parity             1.626e-01  1.070e-01   1.520
## age:gestation:smoke:parity                4.053e-01  2.316e-01   1.750
## height:weight:age:gestation:smoke        -2.898e-05  1.685e-05  -1.720
## height:weight:age:gestation:parity       -5.843e-05  4.584e-05  -1.275
## height:weight:age:smoke:parity           -2.905e-02  1.894e-02  -1.534
## height:weight:gestation:smoke:parity     -2.566e-03  1.650e-03  -1.555
## height:age:gestation:smoke:parity        -6.407e-03  3.599e-03  -1.780
## weight:age:gestation:smoke:parity        -6.624e-03  4.314e-03  -1.536
## height:weight:age:gestation:smoke:parity  1.047e-04  6.646e-05   1.576
##                                          Pr(>|t|)
## (Intercept)                                0.9700
## height                                     0.9841
## weight                                     0.9409
## age                                        0.8941
## gestation                                  0.9325
## smoke                                      0.0917 .
## parity                                     0.1245
## height:weight                              0.9365
## height:age                                 0.9169
## weight:age                                 0.8679
## height:gestation                           0.9387
## weight:gestation                           0.9004
## age:gestation                              0.8562
## height:smoke                               0.0986 .
## weight:smoke                               0.1017
## age:smoke                                  0.0786 .
## gestation:smoke                            0.0828 .
## height:parity                              0.1230
## weight:parity                              0.1896
## age:parity                                 0.1519
## gestation:parity                           0.1069
## smoke:parity                               0.0944 .
## height:weight:age                          0.8738
## height:weight:gestation                    0.8964
## height:age:gestation                       0.8785
## weight:age:gestation                       0.8303
## height:weight:smoke                        0.1058
## height:age:smoke                           0.0845 .
## weight:age:smoke                           0.0901 .
## height:gestation:smoke                     0.0890 .
## weight:gestation:smoke                     0.0926 .
## age:gestation:smoke                        0.0708 .
## height:weight:parity                       0.1863
## height:age:parity                          0.1495
## weight:age:parity                          0.2339
## height:gestation:parity                    0.1055
## weight:gestation:parity                    0.1644
## age:gestation:parity                       0.1329
## height:smoke:parity                        0.0899 .
## weight:smoke:parity                        0.1388
## age:smoke:parity                           0.0888 .
## gestation:smoke:parity                     0.0857 .
## height:weight:age:gestation                0.8360
## height:weight:age:smoke                    0.0943 .
## height:weight:gestation:smoke              0.0964 .
## height:age:gestation:smoke                 0.0761 .
## weight:age:gestation:smoke                 0.0820 .
## height:weight:age:parity                   0.2298
## height:weight:gestation:parity             0.1612
## height:age:gestation:parity                0.1306
## weight:age:gestation:parity                0.2067
## height:weight:smoke:parity                 0.1305
## height:age:smoke:parity                    0.0836 .
## weight:age:smoke:parity                    0.1349
## height:gestation:smoke:parity              0.0811 .
## weight:gestation:smoke:parity              0.1287
## age:gestation:smoke:parity                 0.0804 .
## height:weight:age:gestation:smoke          0.0858 .
## height:weight:age:gestation:parity         0.2027
## height:weight:age:smoke:parity             0.1254
```

```
## height:weight:age:smoke:parity              0.1254
## height:weight:gestation:smoke:parity       0.1201
## height:age:gestation:smoke:parity          0.0753 .
## weight:age:gestation:smoke:parity          0.1249
## height:weight:age:gestation:smoke:parity   0.1153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4401 on 1114 degrees of freedom
## Multiple R-squared:  0.3041, Adjusted R-squared:  0.2648
## F-statistic: 7.728 on 63 and 1114 DF,  p-value: < 2.2e-16
```

Jak widać model jest olbrzymi, jednak większość zmiennych jest nieistotna statystycznie. Jest tak dlatego, że informacja została podzielona pomiędzy mnóstwo predyktorów i interakcji pomiędzy nimi. Niestety funkcja `step` niewiele pomaga.

```
step(model.full)
```

```
## Start:  AIC=-1871.49
## bwt ~ height * weight * age * gestation * smoke * parity
##
##                                          Df Sum of Sq    RSS     AIC
## <none>                                                 215.77 -1871.5
## - height:weight:age:gestation:smoke:parity  1   0.48104 216.25 -1870.9
```

```
##
## Call:
## lm(formula = bwt ~ height * weight * age * gestation * smoke *
##     parity, data = na.omit(babies))
##
## Coefficients:
##                               (Intercept)
##                                 1.261e+01
##                                    height
##                                -1.055e-01
##                                    weight
##                                -4.255e-01
##                                       age
##                                -1.601e+00
##                                 gestation
##                                -1.022e-01
##                                     smoke
##                                -8.156e+02
##                                    parity
##                                -1.900e+03
##                             height:weight
##                                 7.196e-03
##                                height:age
##                                 1.985e-02
##                                weight:age
##                                 3.392e-02
##                           height:gestation
##                                 1.467e-03
##                           weight:gestation
##                                 2.584e-03
##                              age:gestation
##                                 7.845e-03
##                               height:smoke
##                                 1.251e+01
##                               weight:smoke
##                                 1.360e+01
##                                  age:smoke
##                                 3.076e+01
##                            gestation:smoke
##                                 3.018e+00
##                              height:parity
##                                 2.956e+01
##                              weight:parity
##                                 2.823e+01
##                                 age:parity
##                                 7.083e+01
##                           gestation:parity
```

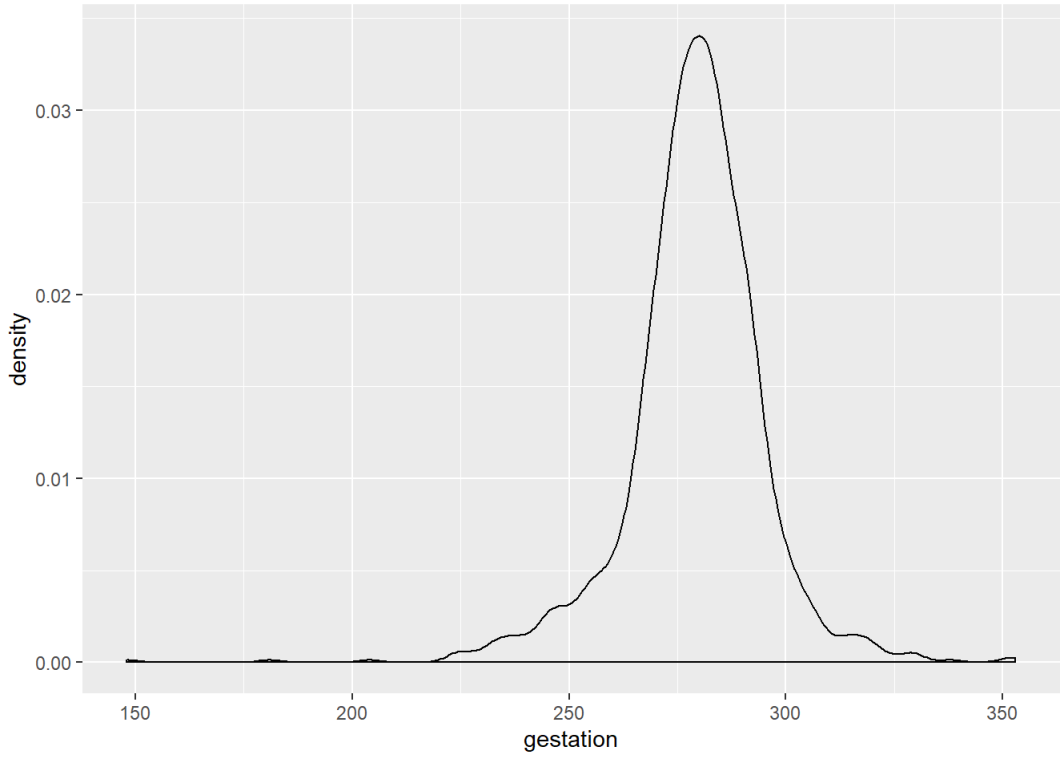
```
##                                6.954e+00
##                             smoke:parity
##                                2.763e+03
##                          height:weight:age
##                                -5.102e-04
##                 height:weight:gestation
##                                -4.234e-05
##                    height:age:gestation
##                                -1.045e-04
##                    weight:age:gestation
##                                -1.572e-04
##                     height:weight:smoke
##                                -2.100e-01
##                        height:age:smoke
##                                -4.734e-01
##                        weight:age:smoke
##                                -5.086e-01
##                  height:gestation:smoke
##                                -4.635e-02
##                  weight:gestation:smoke
##                                -5.009e-02
##                     age:gestation:smoke
##                                -1.139e-01
##                    height:weight:parity
##                                -4.397e-01
##                       height:age:parity
##                                -1.101e+00
##                       weight:age:parity
##                                -1.018e+00
##                  height:gestation:parity
##                                -1.082e-01
##                  weight:gestation:parity
##                                -1.044e-01
##                     age:gestation:parity
##                                -2.587e-01
##                      height:smoke:parity
##                                -4.357e+01
##                      weight:smoke:parity
##                                -4.511e+01
##                         age:smoke:parity
##                                -1.124e+02
##                   gestation:smoke:parity
##                                -9.965e+00
##              height:weight:age:gestation
##                                 2.390e-06
##                  height:weight:age:smoke
##                                 7.855e-03
##           height:weight:gestation:smoke
##                                 7.736e-04
##              height:age:gestation:smoke
##                                 1.754e-03
##              weight:age:gestation:smoke
##                                 1.876e-03
##                height:weight:age:parity
##                                 1.583e-02
##           height:weight:gestation:parity
##                                 1.625e-03
##              height:age:gestation:parity
##                                 4.019e-03
##              weight:age:gestation:parity
##                                 3.762e-03
##               height:weight:smoke:parity
##                                 7.114e-01
##                  height:age:smoke:parity
##                                 1.777e+00
##                  weight:age:smoke:parity
##                                 1.838e+00
##            height:gestation:smoke:parity
##                                 1.572e-01
##            weight:gestation:smoke:parity
##                                 1.626e-01
##               age:gestation:smoke:parity
##                                 4.053e-01
```

```
##        height:weight:age:gestation:smoke
##                               -2.898e-05
##       height:weight:age:gestation:parity
##                               -5.843e-05
##           height:weight:age:smoke:parity
##                               -2.905e-02
##     height:weight:gestation:smoke:parity
##                               -2.566e-03
##        height:age:gestation:smoke:parity
##                               -6.407e-03
##        weight:age:gestation:smoke:parity
##                               -6.624e-03
## height:weight:age:gestation:smoke:parity
##                                1.047e-04
```

Funkcja `step` "dochodzi do wniosku", że nie należy usuwać interakcji najwyższego rzędu i w związku z tym zostaje przy pełnym modelu, co jest absurdem.

Rozważmy sensowniejszy model startowy. Warto włączyć do niego wszystkie dostępne zmienne. A zwłaszcza zmienną `gestation`, o której wiemy, że jest najlepiej skorelowana z wagą dziecka (jest to zrozumiałe, gdyż starsze dzieci generalnie są większe). Proponuję również włączyć interakcję pomiędzy długością trwania ciąży, a paleniem papierosów. Podejrzewamy, że palenie papierosów skraca ciążę, co widać na poniższym wykresie:

```
ggplot(na.omit(babies), aes(fill = smoke, x = gestation)) + geom_density()
```




```
model.start <- lm(bwt~height+weight+gestation+age+smoke+parity+smoke:gestation, na.omit(babies))
summary(model.start)
```

```
##
## Call:
## lm(formula = bwt ~ height + weight + gestation + age + smoke +
##     parity + smoke:gestation, data = na.omit(babies))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58207 -0.28323 -0.00004  0.26174  1.42909
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.2616650  0.3917146  -3.221 0.001313 **
## height           0.0250940  0.0045097   5.564 3.26e-08 ***
## weight           0.0036483  0.0015083   2.419 0.015726 *
## gestation        0.0104376  0.0010087  10.348  < 2e-16 ***
## age             -0.0003154  0.0022396  -0.141 0.888040
## smoke           -1.8584604  0.4726714  -3.932 8.93e-05 ***
## parity          -0.0947181  0.0313016  -3.026 0.002532 **
## gestation:smoke  0.0058390  0.0016936   3.448 0.000586 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4424 on 1170 degrees of freedom
## Multiple R-squared:  0.2614, Adjusted R-squared:  0.257
## F-statistic: 59.15 on 7 and 1170 DF,  p-value: < 2.2e-16
```

Jak widać, zapostulowana interakcja zmiennych `smoke` i `gestation` jest istotna statystycznie. Uprośćmy teraz model przy użyciu AIC.

```
model.start.aic <- step(model.start)
```

```
## Start:  AIC=-1913.29
## bwt ~ height + weight + gestation + age + smoke + parity + smoke:gestation
##
##                   Df Sum of Sq    RSS     AIC
## - age              1    0.0039 229.02 -1915.3
## <none>                         229.02 -1913.3
## - weight           1    1.1452 230.16 -1909.4
## - parity           1    1.7923 230.81 -1906.1
## - gestation:smoke  1    2.3266 231.35 -1903.4
## - height           1    6.0608 235.08 -1884.5
##
## Step:  AIC=-1915.27
## bwt ~ height + weight + gestation + smoke + parity + gestation:smoke
##
##                   Df Sum of Sq    RSS     AIC
## <none>                         229.02 -1915.3
## - weight           1    1.1456 230.17 -1911.4
## - parity           1    1.9401 230.96 -1907.3
## - gestation:smoke  1    2.3243 231.35 -1905.4
## - height           1    6.0780 235.10 -1886.4
```

Jak widać zasadne było usunięcie wieku matki jako nieistotnego predyktora. Pozostałe zmienne okazały się ważne (w tym interakcja zmiennych `smoke` i `gestation`). Ponadto wybrany model okazał się mieć niższe AIC niż model pełny

```
extractAIC(model.full)
```

```
## [1]    64.000 -1871.494
```

```
extractAIC(model.start.aic)
```

```
## [1]     7.000 -1915.266
```

```
model.full <- lm(bwt ~ height*weight*age*gestation*smoke*parity, na.omit(babies))
summary(model.full)
```

```
##
## Call:
```

```
## lm(formula = bwt ~ height * weight * age * gestation * smoke *
##     parity, data = na.omit(babies))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57003 -0.27992  0.00092  0.26172  1.37567
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             1.261e+01  3.348e+02   0.038
## height                                 -1.055e-01  5.293e+00  -0.020
## weight                                 -4.255e-01  5.738e+00  -0.074
## age                                    -1.601e+00  1.203e+01  -0.133
## gestation                              -1.022e-01  1.207e+00  -0.085
## smoke                                  -8.156e+02  4.831e+02  -1.688
## parity                                 -1.900e+03  1.236e+03  -1.537
## height:weight                           7.196e-03  9.037e-02   0.080
## height:age                              1.985e-02  1.901e-01   0.104
## weight:age                              3.392e-02  2.039e-01   0.166
## height:gestation                        1.467e-03  1.907e-02   0.077
## weight:gestation                        2.584e-03  2.065e-02   0.125
## age:gestation                           7.845e-03  4.329e-02   0.181
## height:smoke                            1.251e+01  7.570e+00   1.653
## weight:smoke                            1.360e+01  8.303e+00   1.638
## age:smoke                               3.076e+01  1.747e+01   1.760
## gestation:smoke                         3.018e+00  1.738e+00   1.736
## height:parity                           2.956e+01  1.915e+01   1.543
## weight:parity                           2.823e+01  2.151e+01   1.313
## age:parity                              7.083e+01  4.940e+01   1.434
## gestation:parity                        6.954e+00  4.310e+00   1.613
## smoke:parity                            2.763e+03  1.650e+03   1.674
## height:weight:age                      -5.102e-04  3.211e-03  -0.159
## height:weight:gestation                -4.234e-05  3.250e-04  -0.130
## height:age:gestation                   -1.045e-04  6.838e-04  -0.153
## weight:age:gestation                   -1.572e-04  7.334e-04  -0.214
## height:weight:smoke                    -2.100e-01  1.298e-01  -1.619
## height:age:smoke                       -4.734e-01  2.741e-01  -1.727
## weight:age:smoke                       -5.086e-01  2.998e-01  -1.696
## height:gestation:smoke                 -4.635e-02  2.723e-02  -1.702
## weight:gestation:smoke                 -5.009e-02  2.976e-02  -1.683
## age:gestation:smoke                    -1.139e-01  6.299e-02  -1.808
## height:weight:parity                   -4.397e-01  3.326e-01  -1.322
## height:age:parity                      -1.101e+00  7.637e-01  -1.442
## weight:age:parity                      -1.018e+00  8.545e-01  -1.191
## height:gestation:parity                -1.082e-01  6.676e-02  -1.620
## weight:gestation:parity                -1.044e-01  7.506e-02  -1.391
## age:gestation:parity                   -2.587e-01  1.720e-01  -1.504
## height:smoke:parity                    -4.357e+01  2.566e+01  -1.698
## weight:smoke:parity                    -4.511e+01  3.045e+01  -1.481
## age:smoke:parity                       -1.124e+02  6.600e+01  -1.703
## gestation:smoke:parity                 -9.965e+00  5.794e+00  -1.720
## height:weight:age:gestation             2.390e-06  1.154e-05   0.207
## height:weight:age:smoke                 7.855e-03  4.690e-03   1.675
## height:weight:gestation:smoke           7.736e-04  4.649e-04   1.664
## height:age:gestation:smoke              1.754e-03  9.882e-04   1.775
## weight:age:gestation:smoke              1.876e-03  1.078e-03   1.741
## height:weight:age:parity                1.583e-02  1.317e-02   1.202
## height:weight:gestation:parity          1.625e-03  1.159e-03   1.402
## height:age:gestation:parity             4.019e-03  2.657e-03   1.513
## weight:age:gestation:parity             3.762e-03  2.978e-03   1.263
## height:weight:smoke:parity              7.114e-01  4.701e-01   1.513
## height:age:smoke:parity                 1.777e+00  1.026e+00   1.732
## weight:age:smoke:parity                 1.838e+00  1.228e+00   1.496
## height:gestation:smoke:parity           1.572e-01  9.005e-02   1.746
## weight:gestation:smoke:parity           1.626e-01  1.070e-01   1.520
## age:gestation:smoke:parity              4.053e-01  2.316e-01   1.750
## height:weight:age:gestation:smoke      -2.898e-05  1.685e-05  -1.720
## height:weight:age:gestation:parity     -5.843e-05  4.584e-05  -1.275
## height:weight:age:smoke:parity         -2.905e-02  1.894e-02  -1.534
## height:weight:gestation:smoke:parity   -2.566e-03  1.650e-03  -1.555
## height:age:gestation:smoke:parity      -6.407e-03  3.599e-03  -1.780
## weight:age:gestation:smoke:parity      -6.624e-03  4.314e-03  -1.536
## height:weight:age:gestation:smoke:parity  1.047e-04  6.646e-05   1.576
```

```
##                                         Pr(>|t|)
## (Intercept)                               0.9700
## height                                    0.9841
## weight                                    0.9409
## age                                       0.8941
## gestation                                 0.9325
## smoke                                     0.0917 .
## parity                                    0.1245
## height:weight                             0.9365
## height:age                                0.9169
## weight:age                                0.8679
## height:gestation                          0.9387
## weight:gestation                          0.9004
## age:gestation                             0.8562
## height:smoke                              0.0986 .
## weight:smoke                              0.1017
## age:smoke                                 0.0786 .
## gestation:smoke                           0.0828 .
## height:parity                             0.1230
## weight:parity                             0.1896
## age:parity                                0.1519
## gestation:parity                          0.1069
## smoke:parity                              0.0944 .
## height:weight:age                         0.8738
## height:weight:gestation                   0.8964
## height:age:gestation                      0.8785
## weight:age:gestation                      0.8303
## height:weight:smoke                       0.1058
## height:age:smoke                          0.0845 .
## weight:age:smoke                          0.0901 .
## height:gestation:smoke                    0.0890 .
## weight:gestation:smoke                    0.0926 .
## age:gestation:smoke                       0.0708 .
## height:weight:parity                      0.1863
## height:age:parity                         0.1495
## weight:age:parity                         0.2339
## height:gestation:parity                   0.1055
## weight:gestation:parity                   0.1644
## age:gestation:parity                      0.1329
## height:smoke:parity                       0.0899 .
## weight:smoke:parity                       0.1388
## age:smoke:parity                          0.0888 .
## gestation:smoke:parity                    0.0857 .
## height:weight:age:gestation               0.8360
## height:weight:age:smoke                   0.0943 .
## height:weight:gestation:smoke             0.0964 .
## height:age:gestation:smoke                0.0761 .
## weight:age:gestation:smoke                0.0820 .
## height:weight:age:parity                  0.2298
## height:weight:gestation:parity            0.1612
## height:age:gestation:parity               0.1306
## weight:age:gestation:parity               0.2067
## height:weight:smoke:parity                0.1305
## height:age:smoke:parity                   0.0836 .
## weight:age:smoke:parity                   0.1349
## height:gestation:smoke:parity             0.0811 .
## weight:gestation:smoke:parity             0.1287
## age:gestation:smoke:parity                0.0804 .
## height:weight:age:gestation:smoke         0.0858 .
## height:weight:age:gestation:parity        0.2027
## height:weight:age:smoke:parity            0.1254
## height:weight:gestation:smoke:parity      0.1201
## height:age:gestation:smoke:parity         0.0753 .
## weight:age:gestation:smoke:parity         0.1249
## height:weight:age:gestation:smoke:parity  0.1153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4401 on 1114 degrees of freedom
## Multiple R-squared:  0.3041, Adjusted R-squared:  0.2648
## F-statistic: 7.728 on 63 and 1114 DF,  p-value: < 2.2e-16
```

```
extractAIC(model.start.aic)
```

```
## [1]     7.000 -1915.266
```