Verzani Problem Set

Next are considered the problems from Verzani's book on page 89.

Problem 14.1

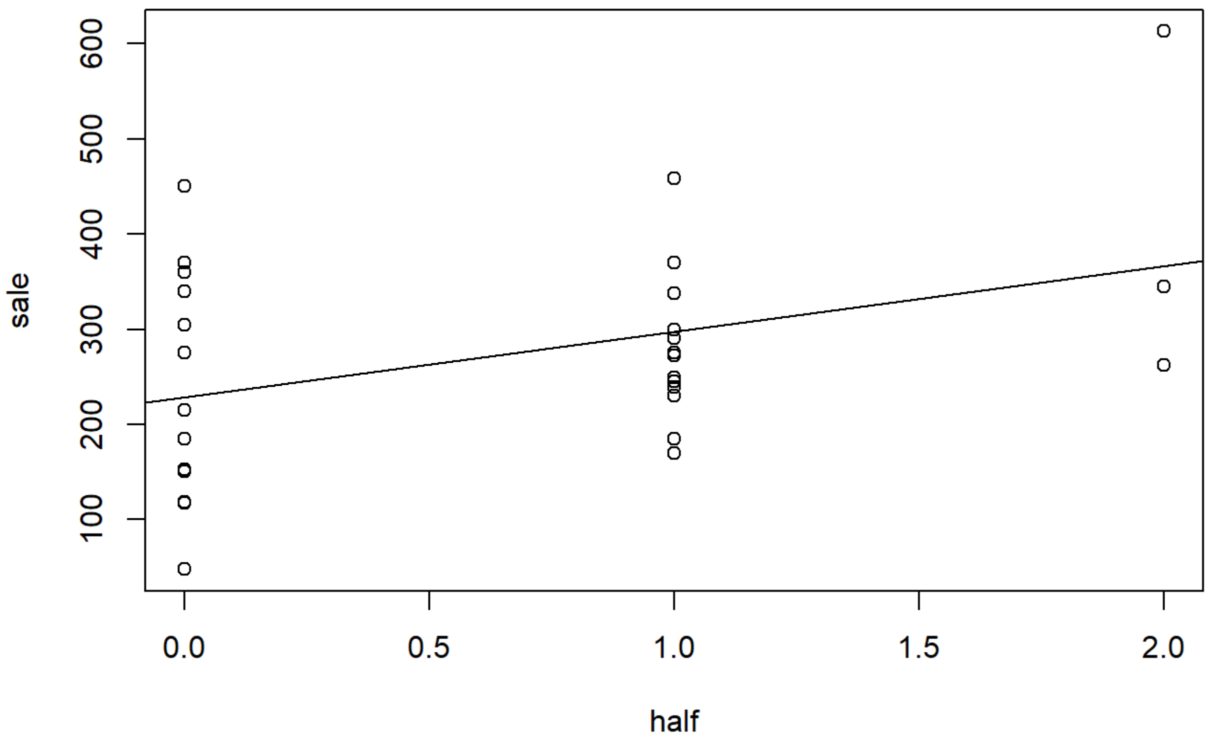
For the `homeprice` data set, what does a half bathroom do for the `sale` price?

Solution

The `homeprice` data set contains information about homes that are sold in a town of New Jersey in the year 2001. We want to figure out what are the appropriate prices in $1000\$$ (denoted by `sale`) for homes.

```
> library(UsingR)
Warning: package 'UsingR' was built under R version 4.0.3
Loading required package: MASS
Loading required package: HistData
Loading required package: Hmisc
Loading required package: lattice
Loading required package: survival
Loading required package: Formula
Loading required package: ggplot2

Attaching package: 'Hmisc'
The following objects are masked from 'package:base':

    format.pval, units

Attaching package: 'UsingR'
The following object is masked from 'package:survival':

    cancer
> head(homeprice)
   list  sale full half bedrooms rooms neighborhood
1  80.0 117.7    1    0        3     6            1
2 151.4 151.0    1    0        4     7            1
3 310.0 300.0    2    1        4     9            3
```

```
4 295.0 275.0     2     1         4     8           3
5 339.0 340.0     2     0         3     7           4
6 337.5 337.5     1     1         4     8           3
> attach(homeprice)
> modelPriceBathroom <- lm(sale ~ half)
> plot(half, sale)
> abline(lm(sale ~ half))
```




```
> summary(modelPriceBathroom)

Call:
lm(formula = sale ~ half)

Residuals:
    Min      1Q  Median      3Q     Max
-180.27  -75.27  -22.34   72.66  246.58

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   228.27      28.78   7.932 1.59e-08 ***
half           69.08      31.00   2.229   0.0344 *
---
```

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

```
Residual standard error: 109.8 on 27 degrees of freedom
Multiple R-squared:  0.1554,	Adjusted R-squared:
0.1241
F-statistic: 4.966 on 1 and 27 DF,  p-value: 0.03436
```

The model is

$$sale = 228.27 + 69.08\, half + \varepsilon$$

One more half bathroom increases the price with $69\,080\$$. In order to compute the $95\ \%$ confidence interval we use our function

```
> myCI <- function(b, SE, t) {
+   b + c(-1,1) * SE * t
+ }
```

In this case first we have to compute

```
> e <- resid(modelPriceBathroom)
> n <- length(e)
> beta1hat <- modelPriceBathroom$coefficients[2];
beta1hat
    half
69.07747
> SSE <- sum(e^2)
> MSE <- SSE / (n-2)
> Seps <- sqrt(MSE)
> SEbeta1 <- Seps / sqrt(sum((full - mean(full))^2));
SEbeta1
[1] 27.63328
> alpha <- 0.05
> t <- qt(1 - alpha/2, n - 2, lower.tail = TRUE)
> myCI(beta1hat, SEbeta1, t)
[1]  12.37866 125.77628
```

Problem 14.2

For the `homeprice` data set, how do the coefficients change if you force the intercept, β_0 to be 0? (Use a 0 or -1 in the model formula notation.) Does it make any sense for this model to have no intercept term?

Solution

First let us see the model with intercept

```
> model.all <- lm(list ~ half + full + bedrooms + rooms +
neighborhood)
> summary(model.all)

Call:
lm(formula = list ~ half + full + bedrooms + rooms +
neighborhood)

Residuals:
    Min      1Q  Median      3Q     Max
-60.788 -28.776   4.351  23.859  62.720

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -144.544     36.026  -4.012 0.000546 ***
half           45.556     12.397   3.675 0.001257 **
full           32.125     13.427   2.392 0.025293 *
bedrooms       18.446     17.197   1.073 0.294572
rooms           7.126     10.033   0.710 0.484661
neighborhood   77.430      9.737   7.952 4.75e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 37.97 on 23 degrees of freedom
Multiple R-squared:  0.9183,    Adjusted R-squared:
0.9006
F-statistic: 51.74 on 5 and 23 DF,  p-value: 9.358e-12
```

Let us see now the coefficients without intercept

```
> model.all <- lm(list ~ 0 + half + full + bedrooms +
rooms + neighborhood)
> summary(model.all)

Call:
lm(formula = list ~ 0 + half + full + bedrooms + rooms +
neighborhood)

Residuals:
     Min       1Q   Median       3Q      Max 
-118.547  -27.898   -0.298   25.814   68.001 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
half            54.05      15.59   3.467   0.0020 ** 
full            36.76      17.07   2.153   0.0416 *  
bedrooms        17.76      21.95   0.809   0.4263    
rooms          -10.40      11.53  -0.902   0.3760    
neighborhood    69.11      12.14   5.691 7.31e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 48.46 on 24 degrees of freedom
Multiple R-squared:  0.9783,    Adjusted R-squared:
0.9738 
F-statistic: 216.3 on 5 and 24 DF,  p-value: < 2.2e-16
```

or

```
> model.all <- lm(list ~ -1 + half + full + bedrooms +
rooms + neighborhood)
> summary(model.all)

Call:
lm(formula = list ~ -1 + half + full + bedrooms + rooms +
neighborhood)

Residuals:
     Min       1Q   Median       3Q      Max 
-118.547  -27.898   -0.298   25.814   68.001 
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
half            54.05      15.59   3.467   0.0020 **
full            36.76      17.07   2.153   0.0416 *
bedrooms        17.76      21.95   0.809   0.4263
rooms          -10.40      11.53  -0.902   0.3760
neighborhood    69.11      12.14   5.691 7.31e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 48.46 on 24 degrees of freedom
Multiple R-squared:  0.9783,    Adjusted R-squared:
0.9738
F-statistic: 216.3 on 5 and 24 DF,  p-value: < 2.2e-16
```

When we compare the adjusted R^2 for the models with and without intercept, we observe that the model without intercept is better.

Let us now improve the model. In the first model the intercept is statistically significant. In the last model the independent variables `bedrooms` and `rooms` are not statistically significant. Therefore, let us now exclude one of them.

```
> model.all <- lm(list ~ half + full + bedrooms +
neighborhood)
> summary(model.all)

Call:
lm(formula = list ~ half + full + bedrooms +
neighborhood)

Residuals:
    Min      1Q  Median      3Q     Max
-57.757 -30.942   4.129  27.084  58.609

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -133.404     32.096  -4.156 0.000355 ***
half           48.094     11.748   4.094 0.000416 ***
full           33.355     13.177   2.531 0.018328 *
bedrooms       28.446      9.775   2.910 0.007675 **
```

```
neighborhood   79.057      9.366   8.441  1.2e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 37.57 on 24 degrees of freedom
Multiple R-squared:  0.9166,    Adjusted R-squared:
0.9026
F-statistic: 65.91 on 4 and 24 DF,  p-value: 1.367e-12
```

All the coefficients in this model are statistically significant.

From practical point of view as far as when we do not buy anything we do not pay anything it is reasonable, however, the intercept to be $\beta_0 = 0$.

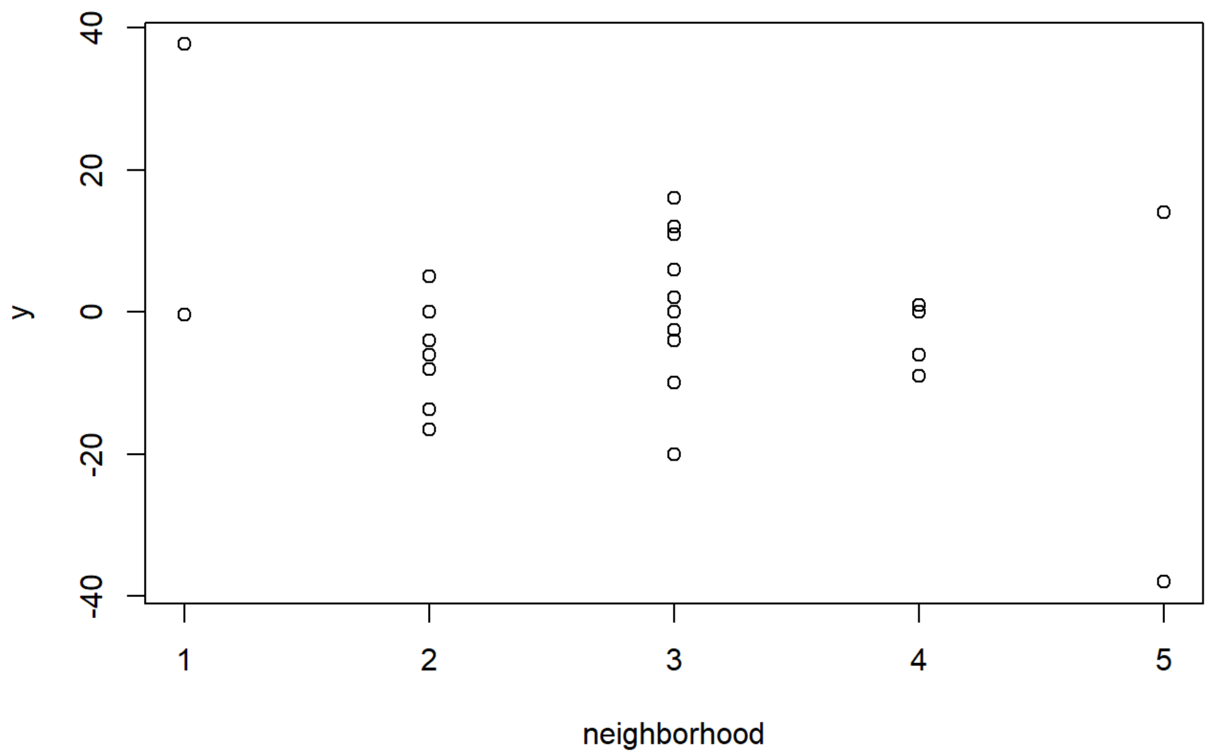
Problem 14.3

For the `homeprice` data set, what is the effect of `neighbourhood` on the difference between `sale` price and `list` price? Do nicer `neighbourhoods` mean it is more likely to have a house go over the asking price?

Solution

```
> y <- sale - list
> model.diff <- lm(y ~ neighbourhood)
> summary(model.diff)

Call:
lm(formula = y ~ neighbourhood)

Residuals:
   Min      1Q Median      3Q     Max
-30.05  -7.50  -0.85   5.80  33.05

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.800      7.435   1.049    0.303
neighbourhood  -3.150      2.428  -1.298    0.205
```

```
Residual standard error: 13 on 27 degrees of freedom
Multiple R-squared:  0.0587,	Adjusted R-squared:  
0.02383 
F-statistic: 1.684 on 1 and 27 DF,  p-value: 0.2054
> plot(neighbourhood, y)
> abline(y ~ neighbourhood)
```



When the points for `neighbourhood` increase with 1, the difference $sale - list$ decreases with $3\ 150\$$. This effect, however, is not statistically significant.

```
> table(neighbourhood)
neighborhood
 1  2  3  4  5 
 2  8 12  5  2 
```

Do nicer `neighbourhood`s mean it is more likely to have a house go over the asking price?

$$H_0 : \mathbb{E}(sale - list | neighbourhood > 3) = 0$$
$$H_A : \mathbb{E}(scale - list | neighbourhood > 3) > 0$$

```
> yall <- sale - list
> y <- yall[neighborhood > 3]
> n <- length(y)
> temp <- (mean(y) - 0) / (sd(y) / sqrt(n)); temp
[1] -0.8663419
> pvalue <- pt(temp, n - 1, lower.tail = FALSE); pvalue
[1] 0.7902036
```

The $p-value = 0.7902036 > \alpha = 0.05$, therefore, we have no evidence to reject H_0. The nicer neighbourhoods does not obligatory mean that it is more likely to have a house go over the asking price.

Problem 14.4

For the `homeprice` data set, is there a relationship between houses which `sell` for more than predicted (a positive residual) and houses which sell for more than asking? (If so, then perhaps the real estate agents aren't pricing the home correctly.)

Solution

Let us first determine the indexes of houses which `sell` for more than asking.

```
> y <- sale - list
> z1 <- which(y > 0); z1
 [1]  1  5  9 10 12 14 18 25 26 29
```

In order to compute the indexes of houses which had been sold for more than predicted (a positive residual) first we build up the general model and then see which residuals are positive.

```
> model.all <- lm(list ~ half + full + bedrooms + rooms +
neighbourhood)
> summary(model.all)

Call:
lm(formula = list ~ half + full + bedrooms + rooms +
neighbourhood)

Residuals:
```

```
    Min      1Q  Median      3Q     Max 
-60.788 -28.776   4.351  23.859  62.720 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -144.544     36.026  -4.012 0.000546 ***
half           45.556     12.397   3.675 0.001257 ** 
full           32.125     13.427   2.392 0.025293 *  
bedrooms       18.446     17.197   1.073 0.294572    
rooms           7.126     10.033   0.710 0.484661    
neighborhood   77.430      9.737   7.952 4.75e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 37.97 on 23 degrees of freedom
Multiple R-squared:  0.9183,	Adjusted R-squared:
0.9006 
F-statistic: 51.74 on 5 and 23 DF,  p-value: 9.358e-12
```

We exclude the statistically insignificant variables.

```
> model.my <- lm(list ~ half + full +  rooms +
neighbourhood)
> summary(model.my)

Call:
lm(formula = list ~ half + full + rooms + neighborhood)

Residuals:
   Min     1Q Median     3Q    Max 
-65.00 -24.78   4.55  22.91  75.70 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -144.162     36.137  -3.989 0.000541 ***
half           44.618     12.405   3.597 0.001449 ** 
full           33.085     13.439   2.462 0.021392 *  
rooms          15.936      5.780   2.757 0.010972 *  
neighbourhood   75.223      9.547   7.879 4.13e-08 ***
---
```

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

Residual standard error: 38.09 on 24 degrees of freedom
Multiple R-squared:  0.9143,    Adjusted R-squared:
0.9
F-statistic: 63.98 on 4 and 24 DF,  p-value: 1.889e-12
```

We compute the indexes of houses which had been sold for more than predicted (a positive residual)

```
> e <- resid(model.my)
> z2 <- which(e > 0); z2
 1  2  5  6  7  8  9 11 12 14 15 17 19 20 26
 1  2  5  6  7  8  9 11 12 14 15 17 19 20 26
```

and test if the distributions and more precisely the means μ_1 and μ_2 of the populations which correspond to the sample z_1 and z_2 coincide.

$H_0 : \mu_1 \neq \mu_2$
$H_A : \mu_1 \neq \mu_2$

```
> var(z1)
[1] 88.1
> var(z2)
[1] 49.98095
```

First we test

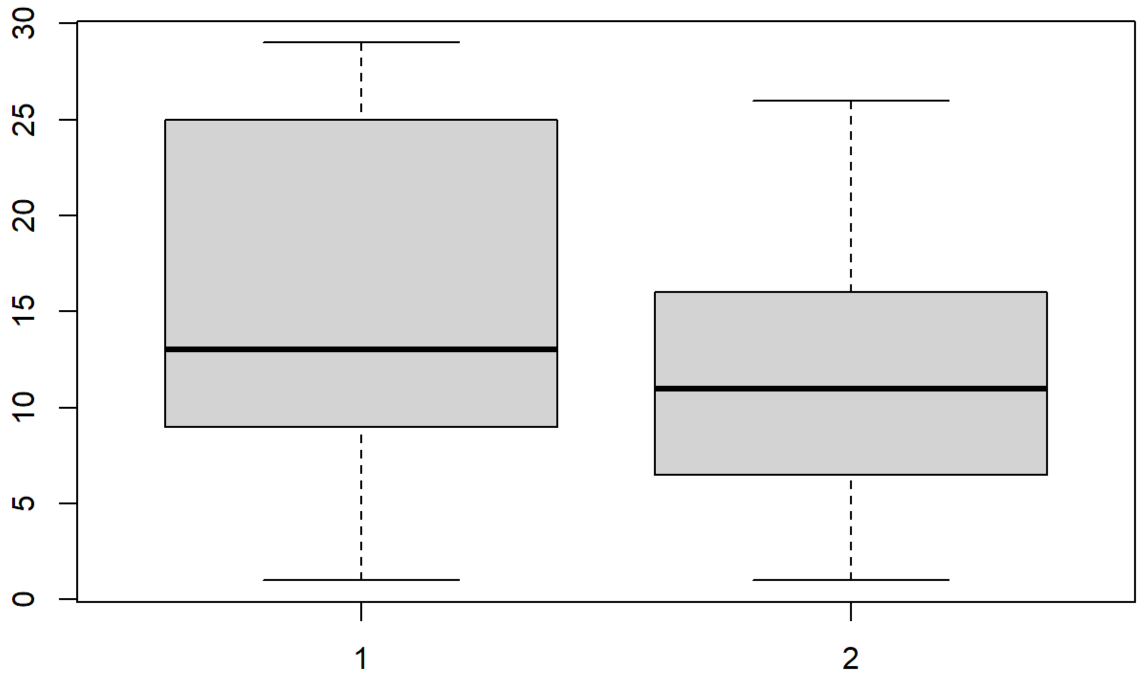
$H_0 : \sigma_1 = \sigma_2$
$H_A : \sigma_1 \neq \sigma_2$

```
> var.test(z1, z2, alternative = "two.sided")

    F test to compare two variances

data:  z1 and z2
F = 1.7627, num df = 9, denom df = 14, p-value = 0.3296
alternative hypothesis: true ratio of variances is not
equal to 1
95 percent confidence interval:
 0.5492386 6.6945426
```

```
sample estimates:
ratio of variances
          1.762671
```

The $p-value = 0.3296 > 0.05 = \alpha$, so we have no evidence to reject H_0.

```
> boxplot(z1, z2)
```



and make the `t-test`

```
> t.test(z1, z2, alternative = "two.sided", var.equal =
TRUE)

    Two Sample t-test

data:  z1 and z2
t = 1.0439, df = 23, p-value = 0.3074
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -3.370061 10.236728
```

```
sample estimates:
mean of x mean of y 
 14.90000  11.46667
```

The $p-value = 0.3074 > 0.05 = \alpha$ therefore we have no evidence to reject H_0, so both population of houses coincide.

Problem 14.5

For the `babies` data set, do a multiple regression of birthweight(`wt`) with regressors the mothers `age`, weight `wt1` and height `ht`. What is the value of R^2? What are the coefficients? Do any variables appear to be 0?

Solution

```
> head(babies)
   id plurality outcome date gestation sex  wt parity race
age ed ht wt1 drace
1  15         5       1 1411       284   1 120      1    8
27  5 62 100     8
2  20         5       1 1499       282   1 113      2    0
33  5 64 135     0
3  58         5       1 1576       279   1 128      1    0
28  2 64 115     5
4  61         5       1 1504       999   1 123      2    0
36  5 69 190     3
5  72         5       1 1425       282   1 108      1    0
23  5 67 125     0
6 100         5       1 1673       286   1 136      4    0
25  2 62  93     3
  dage ded dht dwt marital inc smoke time number
1   31   5  65 110       1   1     0    0      0
2   38   5  70 148       1   4     0    0      0
3   32   1  99 999       1   2     1    1      1
4   43   4  68 197       1   8     3    5      5
5   24   5  99 999       1   1     1    1      5
6   28   2  64 130       1   4     2    2      2
> ls(babies)
 [1] "age"        "dage"       "date"       "ded"       
"dht"        "drace"
```

```
 [7] "dwt"         "ed"          "gestation"  "ht"          "id"
"inc"
[13] "marital"     "number"      "outcome"    "parity"
"plurality"   "race"
[19] "sex"         "smoke"       "time"       "wt"
"wt1"
> attach(babies)
> n <- length(wt); n
[1] 1236
```

This data frame contains $1\,236$ observations. We need the following ones

`wt` - birth weight in ounces (where `999` means unknown)

`ht` - mother's height in inches to the last completed inch (where `99` means unknown)

`age` - mother's age in years at termination of pregnancy, (where `99` means unknown)

`wt1` - mother pregnancy weight in pounds, (where `999` means unknown)

Let us build up the multiple regression model

```
> data <- babies[wt != 999 && ht != 99 && age != 99 &&
wt1 != 999, ]
> View(data)
> model.all <- lm(data$wt ~ data$age + data$wt1 +
data$ht)
> summary(model.all)

Call:
lm(formula = data$wt ~ data$age + data$wt1 + data$ht)

Residuals:
    Min      1Q  Median      3Q     Max
-65.360 -11.304   0.421  11.425  56.682

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 84.456722   7.826997  10.790  < 2e-16 ***
data$age     0.069360   0.079916   0.868    0.386
```

```
data$wt1      -0.005697   0.004361  -1.306     0.192    
data$ht        0.527277   0.122477   4.305   1.8e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 18.1 on 1232 degrees of freedom
Multiple R-squared:  0.01765,   Adjusted R-squared:
0.01526
F-statistic: 7.378 on 3 and 1232 DF,  p-value: 6.69e-05
```

The model is

$$wt = 84.456722 + 0.069360\, age - 0.005697\, wt1 + 0.527277\, ht$$

The variables `age` and `wt1` are not statistically significant.
Adjusted $R^2 = 0.01526$ is small therefore the model is not quite good.