Simple regression model

Structure of the model

A simple regression can be written with two β parameters, a slope β_1 and intercept β_0:

$$Y = \beta_0 + \beta_1 X + e$$

where:

- Y is a vector of observed values
- β_0 is a parameter representing the intercept of the regression line.
- X is a continuous predictor.
- β_1 is a parameter representing the slope of the regression line.
- e is a vector of independent, identically distributed $N(0, \sigma_e)$ random variables
- σ_e is the standard deviation of the error or residual terms e, which are assumed to have mean zero.

Example: Predicting weight by height

Read the data:

```
df    =  read.table("body.dat.txt")        #read the body measurement data
str(df)

## 'data.frame': 507 obs. of  25 variables:
##  $ V1 : num  42.9 43.7 40.1 44.3 42.5 43.3 43.5 44.4 43.5 42 ...
##  $ V2 : num  26 28.5 28.2 29.9 29.9 27 30 29.8 26.5 28 ...
##  $ V3 : num  31.5 33.5 33.3 34 34 31.5 34 33.2 32.1 34 ...
##  $ V4 : num  17.7 16.9 20.9 18.4 21.5 19.6 21.9 21.8 15.5 22.5 ...
##  $ V5 : num  28 30.8 31.7 28.2 29.4 31.3 31.7 28.8 27.5 28 ...
##  $ V6 : num  13.1 14 13.9 13.9 15.2 14 16.1 15.1 14.1 15.6 ...
##  $ V7 : num  10.4 11.8 10.9 11.2 11.6 11.5 12.5 11.9 11.2 12 ...
##  $ V8 : num  18.8 20.6 19.7 20.9 20.7 18.8 20.8 21 18.9 21.1 ...
##  $ V9 : num  14.1 15.1 14.1 15 14.9 13.9 15.6 14.6 13.2 15 ...
##  $ V10: num  106 110 115 104 108 ...
##  $ V11: num  89.5 97 97.5 97 97.5 ...
##  $ V12: num  71.5 79 83.2 77.8 80 82.5 82 76.8 68.5 77.5 ...
##  $ V13: num  74.5 86.5 82.9 78.8 82.5 80.1 84 80.5 69 81.5 ...
##  $ V14: num  93.5 94.8 95 94 98.5 95.3 101 98 89.5 99.8 ...
##  $ V15: num  51.5 51.5 57.3 53 55.4 57.5 60.9 56 50 59.8 ...
```

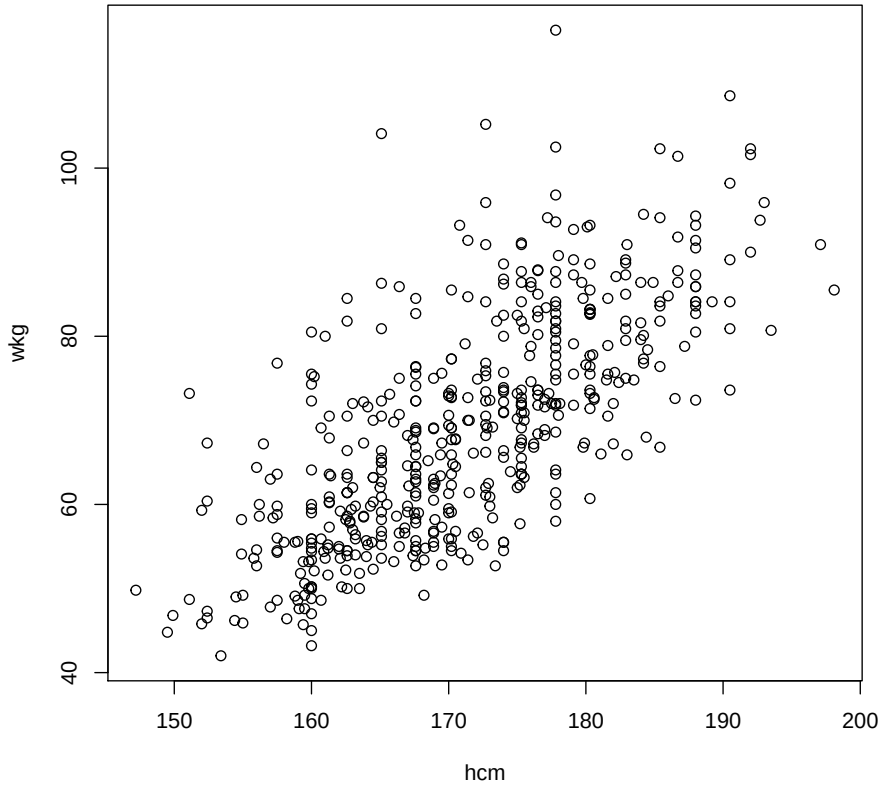
```
##  $ V16: num  32.5 34.4 33.4 31 32 33 42.4 34.1 33 36.5 ...
##  $ V17: num  26 28 28.8 26.2 28.4 28 32.3 28 26 29.2 ...
##  $ V18: num  34.5 36.5 37 37 37.7 36.6 40.1 39.2 35.5 38.3 ...
##  $ V19: num  36.5 37.5 37.3 34.8 38.6 36.1 40.3 36.7 35 38.6 ...
##  $ V20: num  23.5 24.5 21.9 23 24.4 23.5 23.6 22.5 22 22.2 ...
##  $ V21: num  16.5 17 16.9 16.6 18 16.9 18.8 18 16.5 16.9 ...
##  $ V22: num  21 23 28 23 22 21 26 27 23 21 ...
##  $ V23: num  65.6 71.8 80.7 72.6 78.8 74.8 86.4 78.4 62 81.6 ...
##  $ V24: num  174 175 194 186 187 ...
##  $ V25: int  1 1 1 1 1 1 1 1 1 1 ...

wkg   =  df$V23                              #weight in kg

hcm   =  df$V24                              #height in cm
```

Check out the data with a scatter plot

```
plot(wkg~hcm)
```



Now run the regression model.

```
lm1 = lm(wkg ~ hcm)

summary(lm1)

##
## Call:
## lm(formula = wkg ~ hcm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.743  -6.402  -1.231   5.059  41.103
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -105.01125    7.53941  -13.93   <2e-16 ***
## hcm            1.01762    0.04399   23.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.308 on 505 degrees of freedom
## Multiple R-squared:  0.5145,Adjusted R-squared:  0.5136
## F-statistic: 535.2 on 1 and 505 DF,  p-value: < 2.2e-16
```

This is the standard regression model summary.

The regression line has a slope of 1.02 and an intercept of -105.

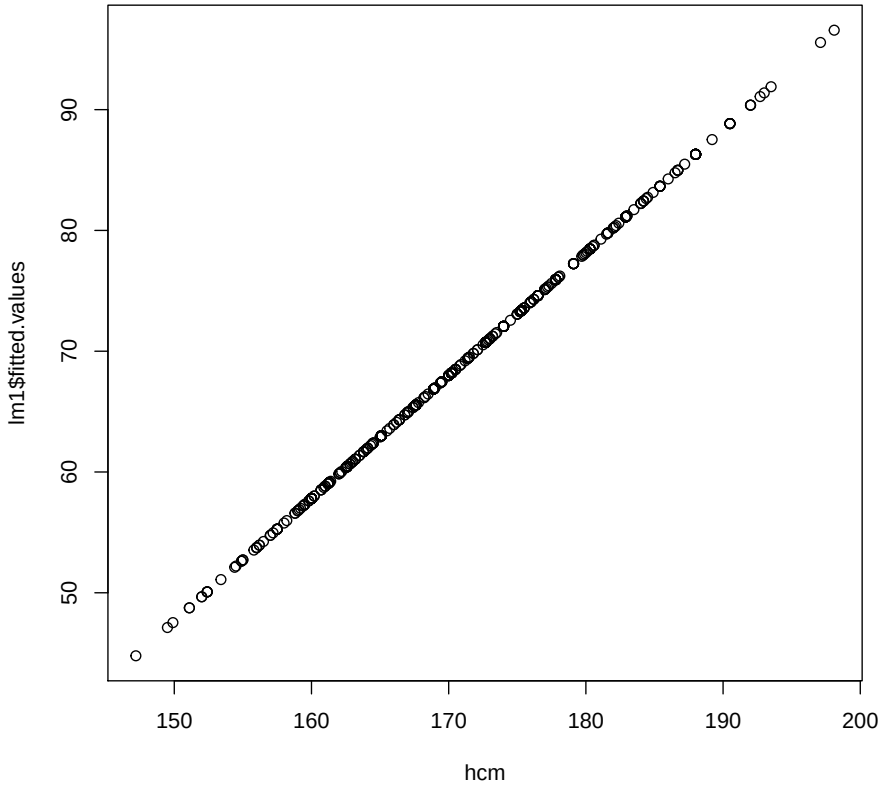
The interpretation in this case is that for each 1 cm of height, the expected weight increases by 1.02kg.

The intercept is the expected weight of a person zero cm tall, which is not meaningful in this setting.

The column `Pr(>|t|)` contains the p-value. A small p-value, say below .05, indicates that the data we observed is not very likely if in fact this parameter is zero.

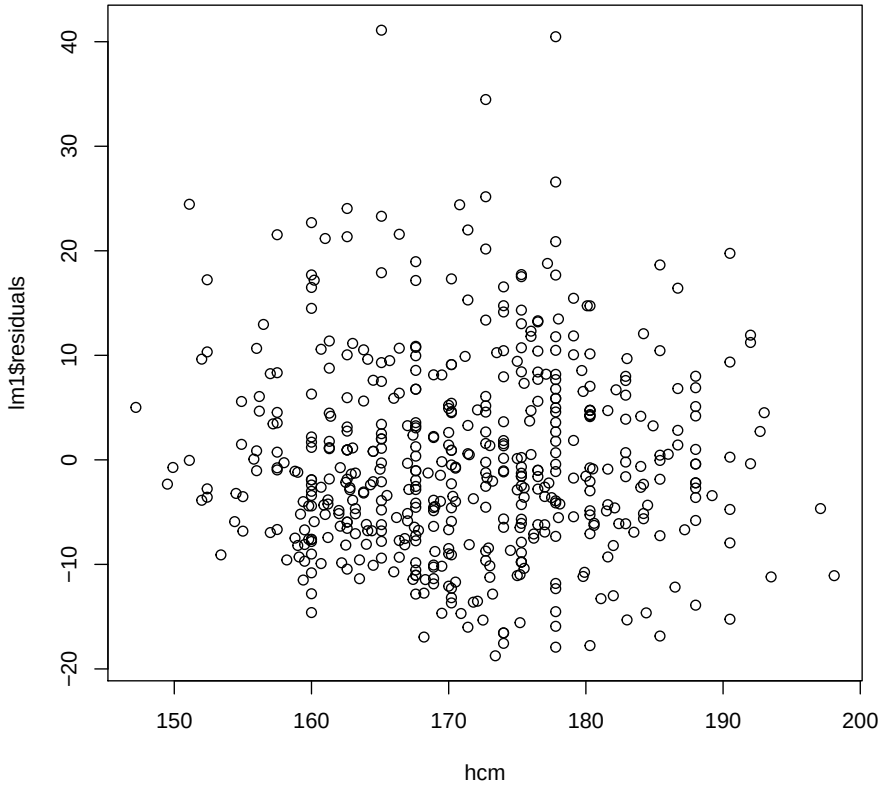
Like any linear model, this one has fitted values. We can examine them by plotting against height:

```
plot(lm1$fitted.values~hcm)
```



We should also examine the residuals for any systematic patterns:

```
plot(lm1$residuals~hcm)
```


lm1$residuals
hcm
40
30
20
10
0
−10
−20
150
160
170
180
190
200