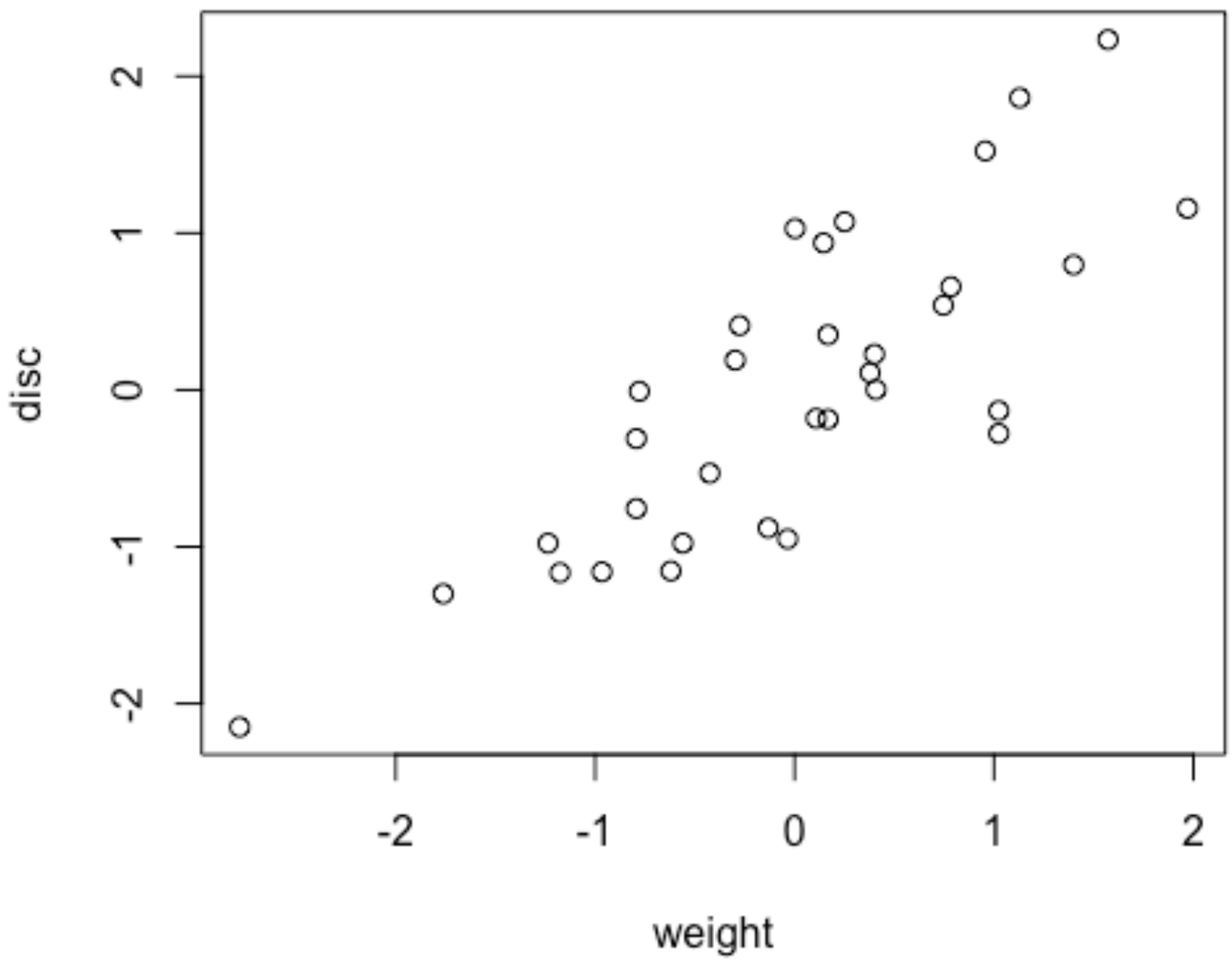
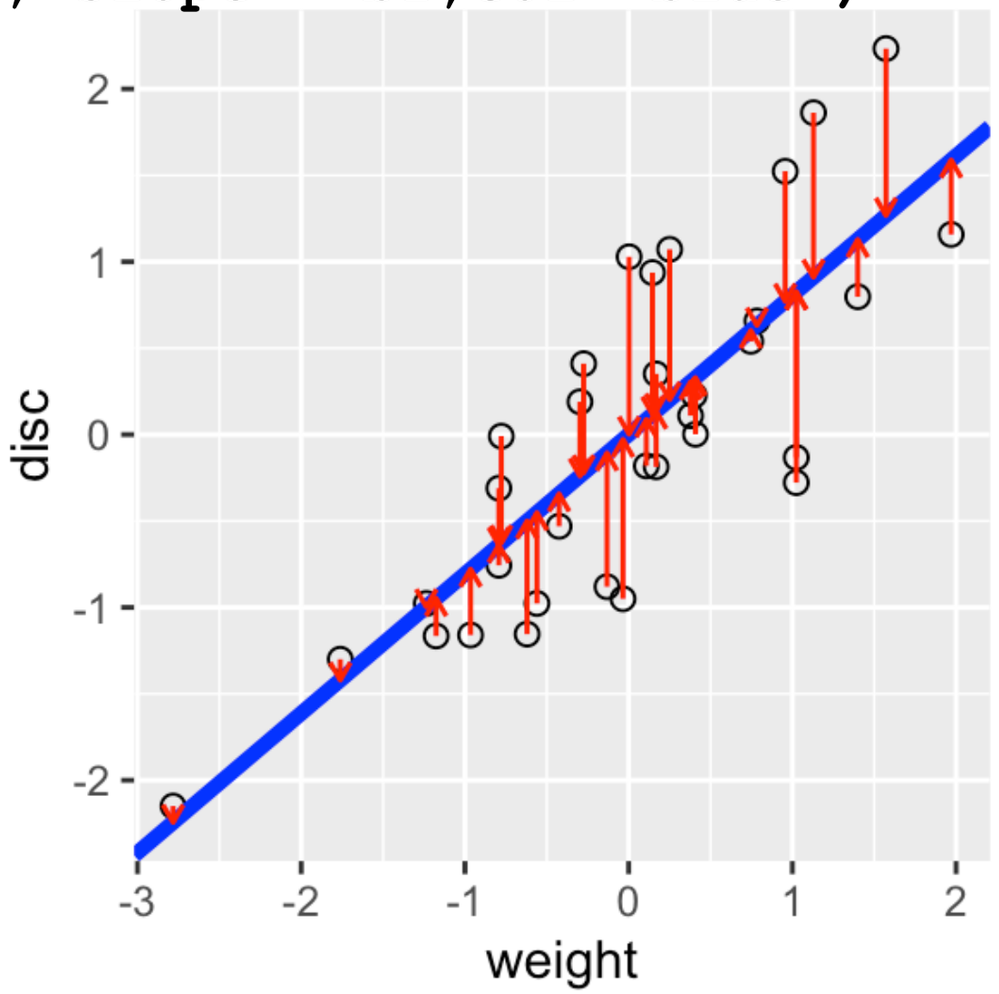
linear regression

```
reg1=lm(disc ~ weight, data=athletes)
a1 = reg1$coefficients[1] # intercept
b1 = reg1$coefficients[2] # slope
ath_gg = ggplot(athletes, aes(x = weight, y = disc)) +
           geom_point(size = 2, shape = 21)
ath_gg + geom_abline(intercept = a1, slope = b1,col="blue”)
```





Regression for prediction

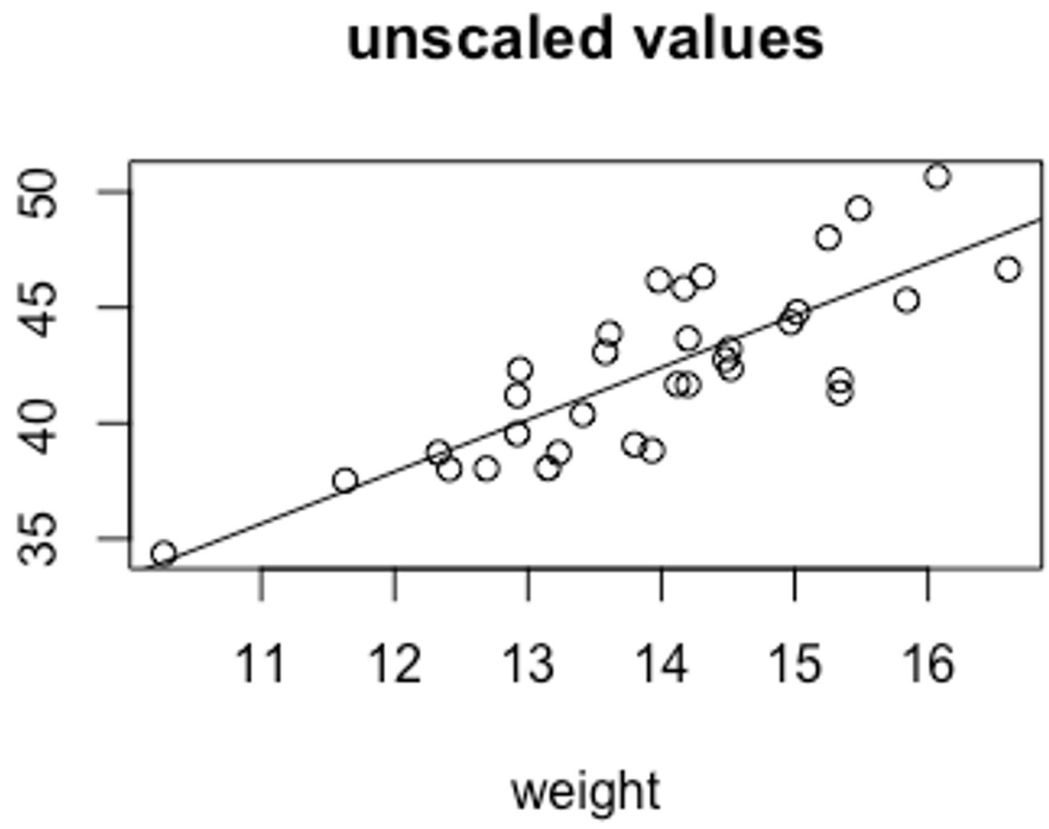
Can we use regression to predict the average discus throw length of someone who can throw the weight 11 meters?

```
plot(weight,disc,main="unscaled values")
abline(lm(disc ~ weight))
```

```
lm(disc ~ weight)

Call:
lm(formula = disc ~ weight)

Coefficients:
(Intercept)       weight
     10.887        2.251
```

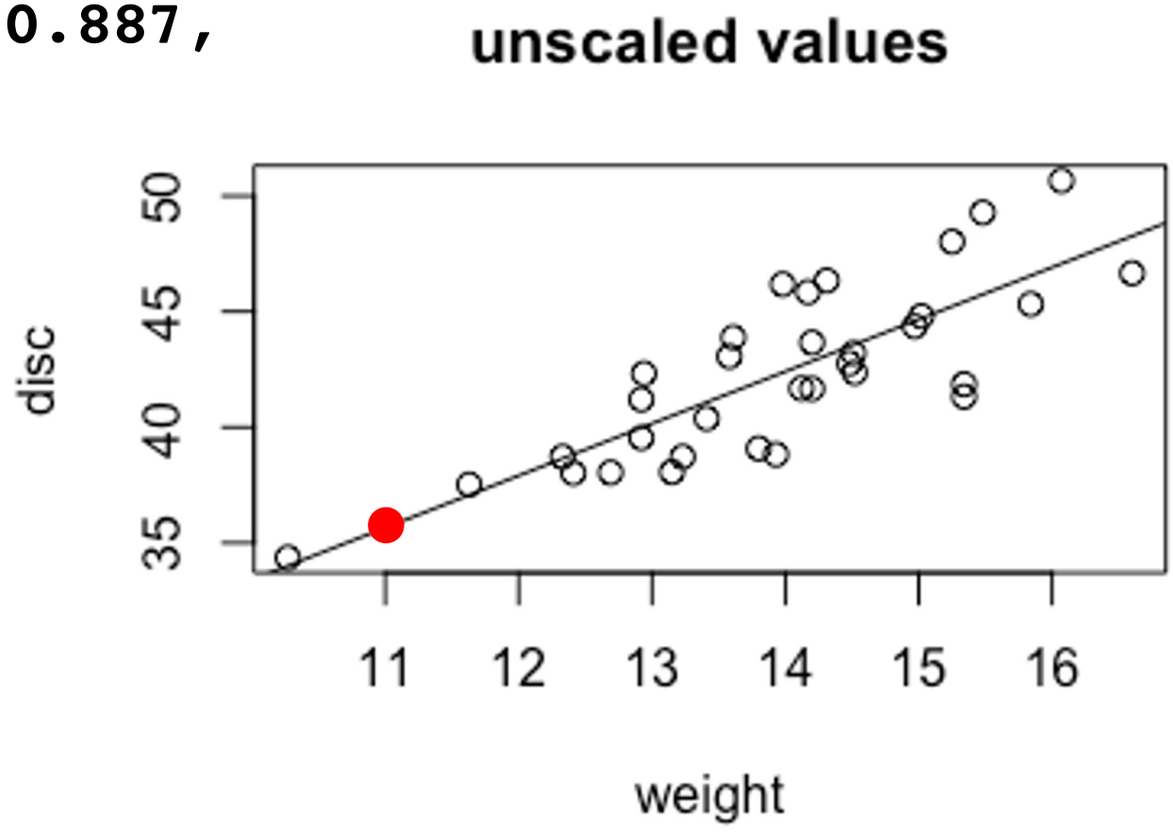


Regression for prediction

Can we use regression to predict the average discus throw length of someone who can throw the weight 11 meters?

y=ax+b, a=2.251, b=10.887

```
points(11, 2.251 * 11 + 10.887,
col="red",pch="16")
```



Multiple linear regression

Can generalize to a linear function of multiple variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots$$

The model is the same, but there are additional terms.

Works the same way, but harder to plot in 2-D.
Still aims to minimize RMSD on the training data:

$$\sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}}$$

```
getRmsd <- function(linmod, truth){  # truth = training data y
  sqrt(sum ( (linmod$fitted.values - truth) **2)/ length(truth))
}
```

Exercise (background)

Longley economic data set

A data frame with 7 economic variables, observed yearly from 1947 to 1962 (*n=16*).

```
GNP:            Gross National Product.
GNP.deflator:   GNP implicit price deflator (1954=100)
Unemployed:     number of unemployed.
Armed.Forces:   number of people in the armed forces.
Population:     population ≥ 14 years of age.
Year:           the year (time).
Employed:       number of people employed.
```

J. W. Longley (1967) An appraisal of least-squares programs from the point of view of the user.*Journal of the American Statistical Association* **62**, 819–841.

linear regression exercise

Load the Longley economic data from the rds file on today's date on the Schedule.

Generate a scatter plot using base-R graphics comparing the number of people `Employed` each year to the `GNP` in the corresponding year. How well do you expect to be able to predict GNP as a function of employment using linear regression?

Using `lm(),` predict GNP (the dependent, or response, variable) from Employed (the independent variable). Use `abline` to add the regression line to the plot.

What is your predicted GNP for a year with employment at 62.2 million? Use the slope and intercept of the regression line to predict. Add this point to your plot, and send me the plot and GNP prediction in Piazza.

multiple linear regression exercise

Now, create a different linear model, also using `lm()`, to predict GNP (the dependent, or response, variable) from three independent variables: Employed, Year, and Population.

To compare the models, you'll need to compute RMSD for each. The following code should do that:

```
# truth = training data y
getRmsd <- function(linmod, truth){
  sqrt(sum ( (linmod$fitted.values - truth) **2)/
               length(truth))
}
```

What is the RMSD you get on the training data for the original model? For the new multivariate one? Which one is better?