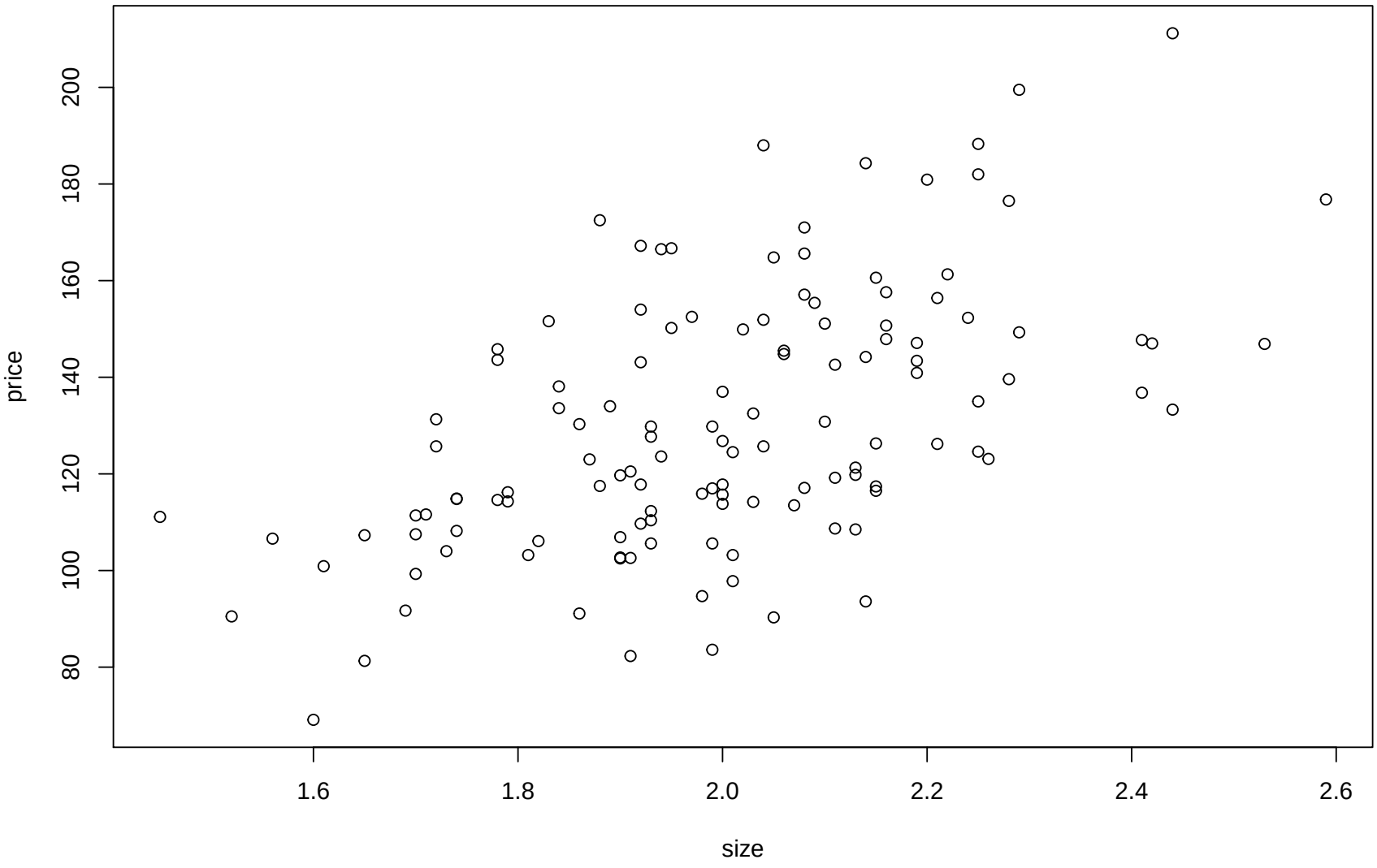
Capturing Relationships with Linear Regression

Rob McCulloch

9/21/2017

Predicting with Simple Linear Regression

Each observation corresponds to a recently sold house.
x=size, y=price.



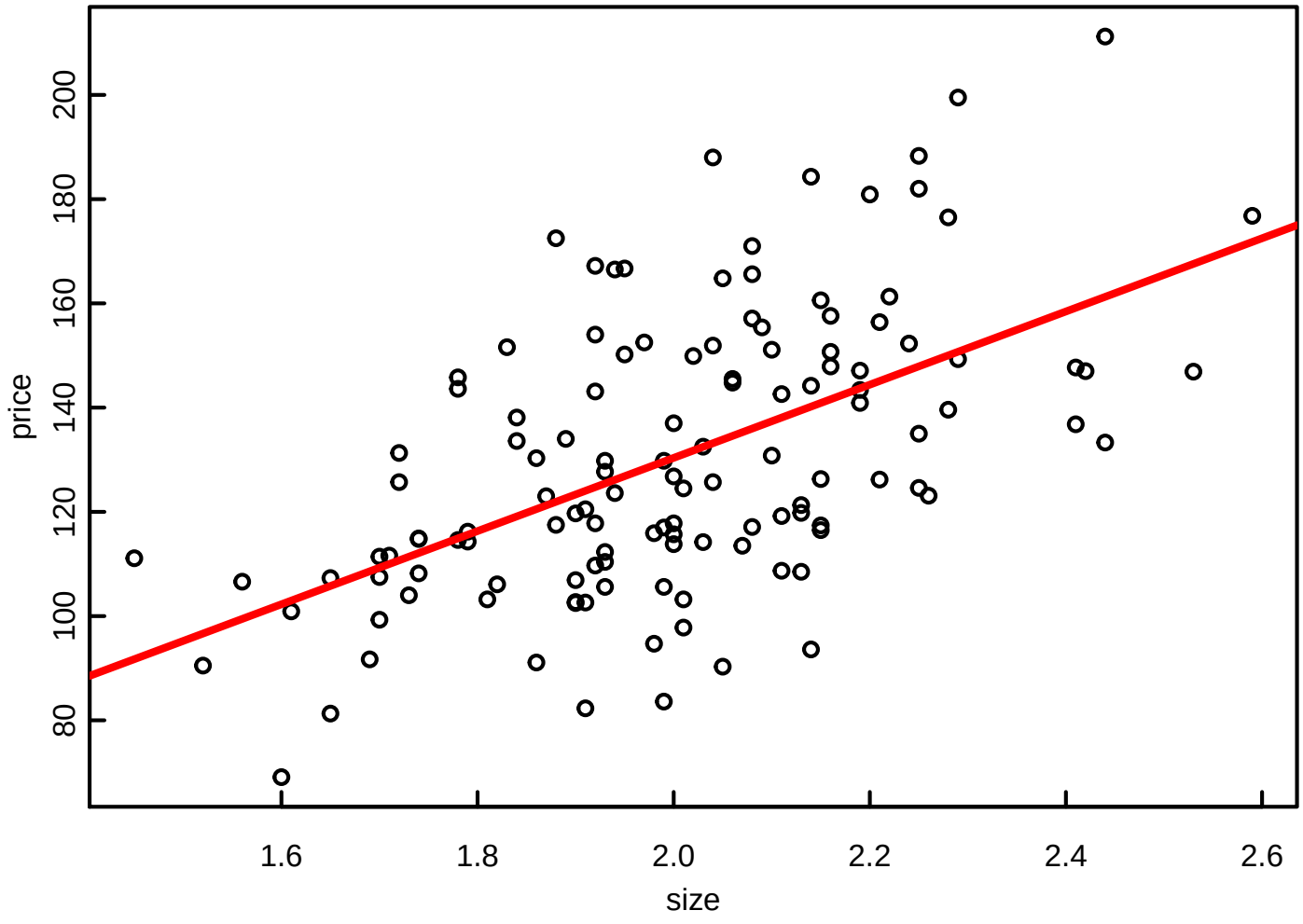
How do you make a prediction????!!!

Well, there are lot's of ways to make the prediction!!

One very nice and time honored approach is to draw a line through the data and use the fitted line to make the prediction.

Linear regression chooses a line for us.

Given the linear pattern, we can make a prediction by drawing a line through the data



Here is the *regression output:*

```
##
## Call:
## lm(formula = price ~ size, data = hdf)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46.59 -16.64  -1.61  15.12  54.83
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -10.091     18.966  -0.532    0.596
## size          70.226      9.426   7.450  1.3e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.48 on 126 degrees of freedom
## Multiple R-squared:  0.3058, Adjusted R-squared:  0.3003
## F-statistic:  55.5 on 1 and 126 DF,  p-value: 1.302e-11
```

We will learn what all the other numbers mean but the key ones for us now are the intercept and slope of the fitted line.

The line regression has fit to the data is:

$$price = -10.091 + 70.226\, size$$

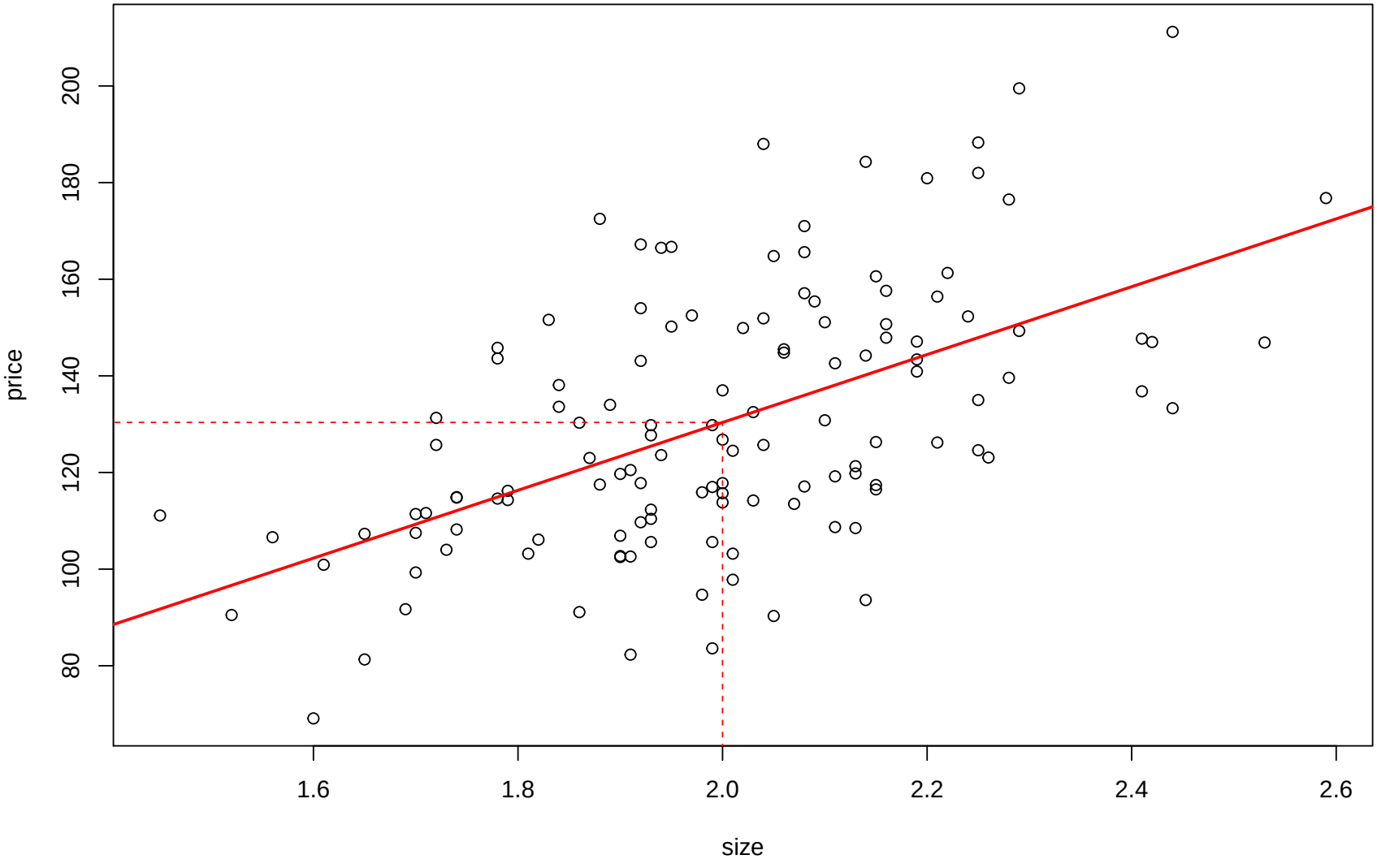
Suppose we know the size of a house is 2.0.

Our prediction for the price is:

$$\widehat{price} = -10.091 + 70.226 * 2 = 130.361$$

We often use the *hat* notation to denote a guess for something we are not sure about.

Plugging into the equation of the line:



The General Notation

In general our notation is:

y: the dependent variable.

x: the independent variable.

How does y depend on x:

$$y = a + b\,x$$

Given data, $\{x_i, y_i\}$ regression chooses values for a and b.

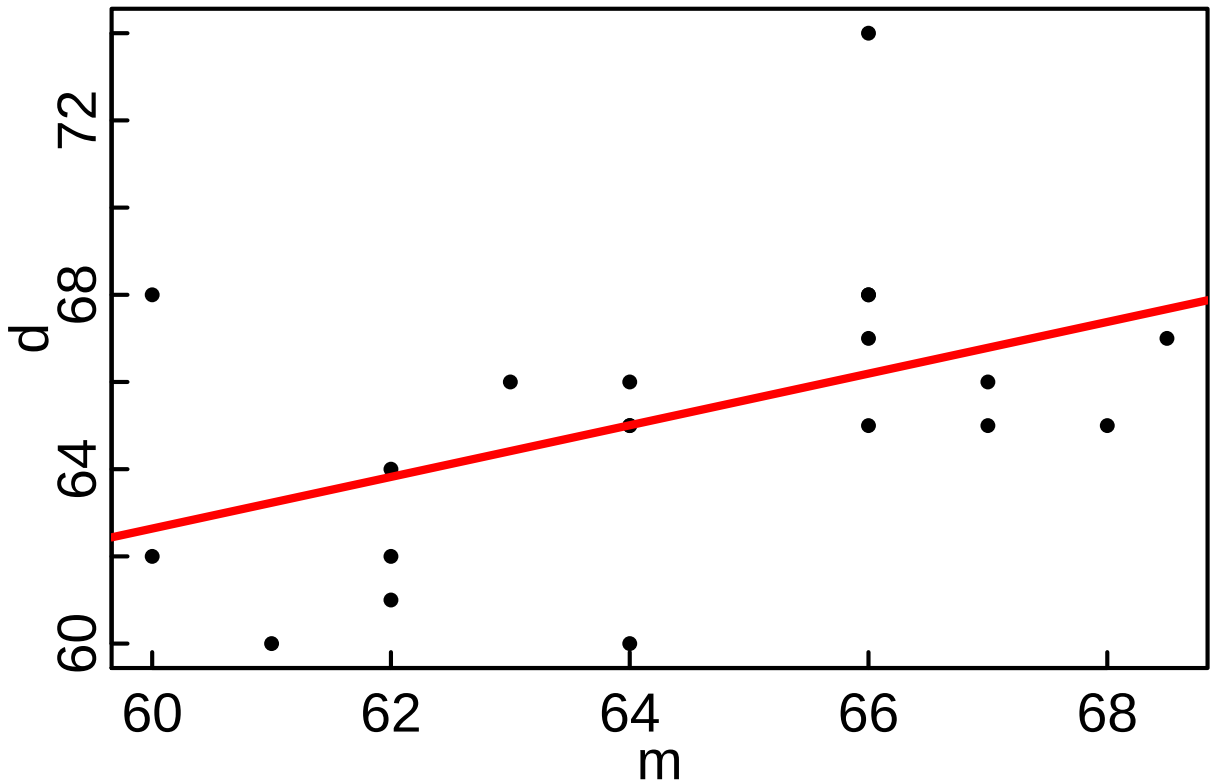
Given a particular value of x_p for the independent variable our prediction is

$$\hat{y} = a + b\,x_p$$

In our house price example, we might say y *is price* and x *is size* or y=*price*, x=*size*.

Mother/Daughter Heights

"y=d": height of daughter
"x=m": height of mother



a: 27.11; b: 0.59

$$d = 27.11 + 0.59\, m$$

Formulas for Least Squares Intercept and Slope

Things can get complicated in predictive modeling, but there are nice simple formulas for a and b which are worth taking a look at.

First of all, *how does regression pick a good line* given the data?

For any choice of a and b we let

$$\hat{y}_i = a + b\, x_i.$$

and

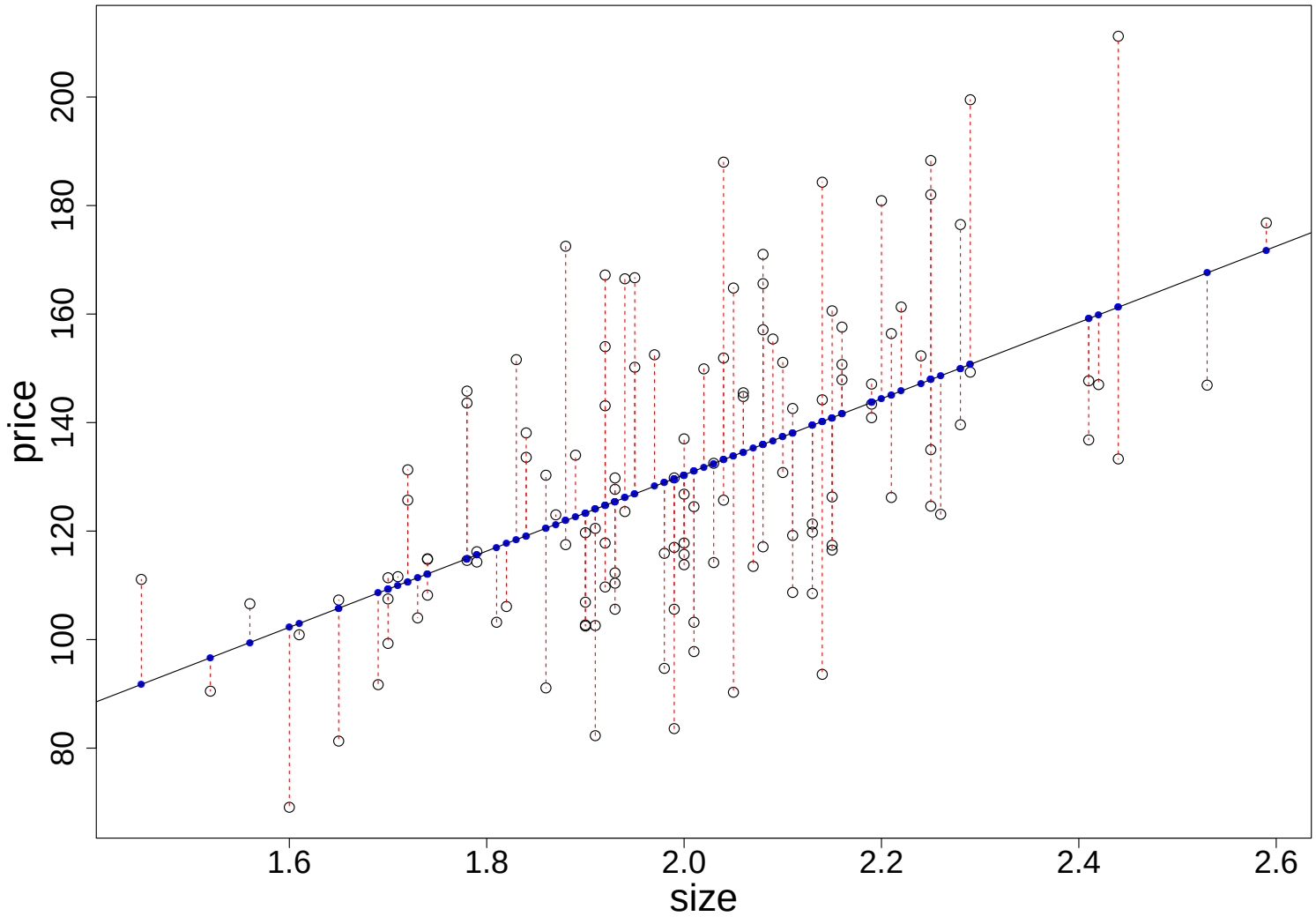
$$e_i = y_i - \hat{y}_i$$

The $\hat{y}_i$ are called the *fitted values*.
The e_i are called the *residuals*.

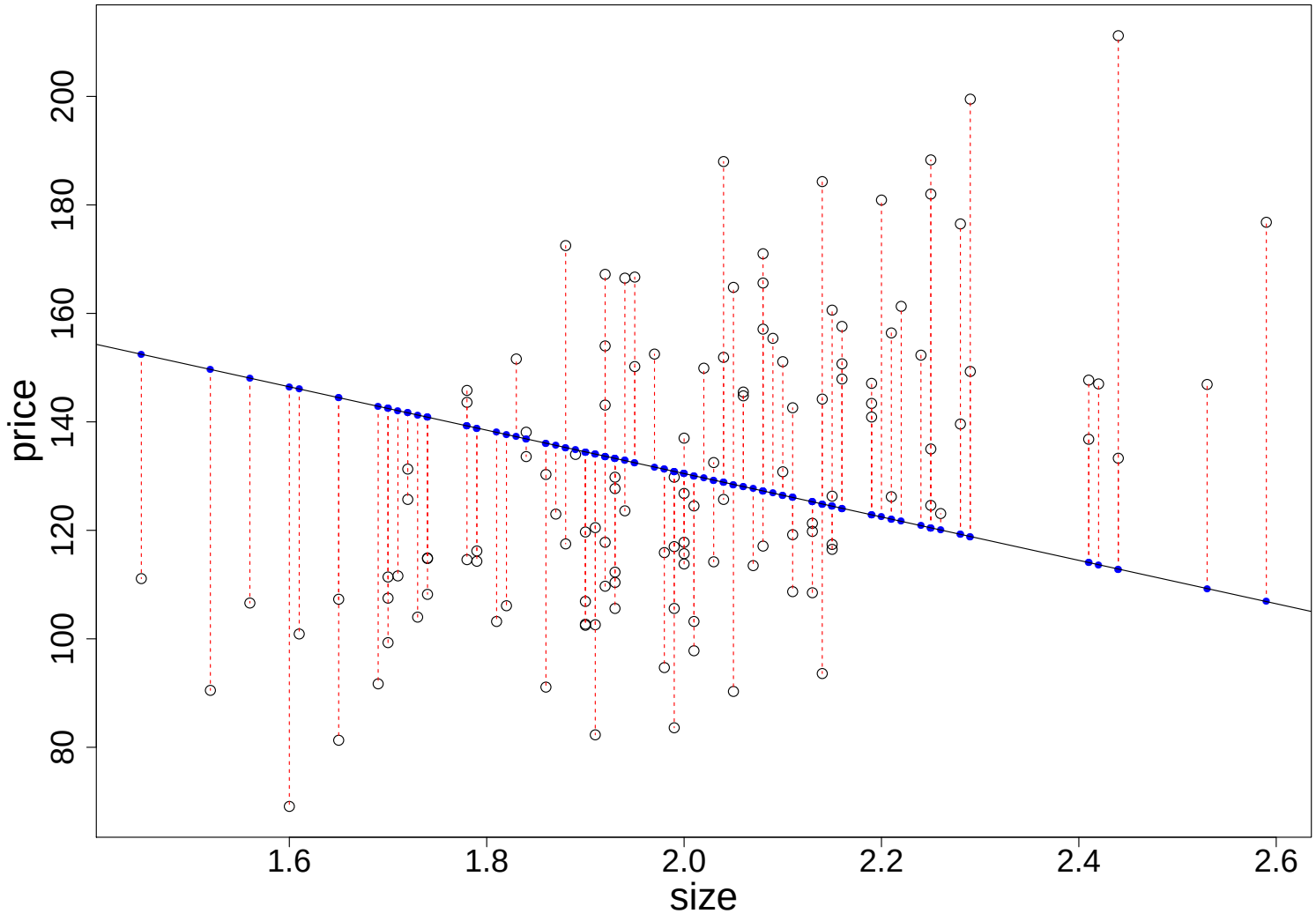
So, the fits are your predictions for y at the observed x_i.

But, you observed y_i so you can compute the error y-fit, the residual.

Fitted values and residuals for the line chosen by regression.
Blue dots on line are fits. Red lines are the residuals.



Fitted values and residuals for a different line.



How do the residuals from the regression line compare with the residuals from this line?

Which line is better ?

Regression chooses the line which make the residual sum of squares as small as possible:

$$\begin{aligned}\underset{a,b}{\text{minimize}} \quad & \sum_{i=1}^{n}(y_i - a - b\,x_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} e_i^2\end{aligned}$$

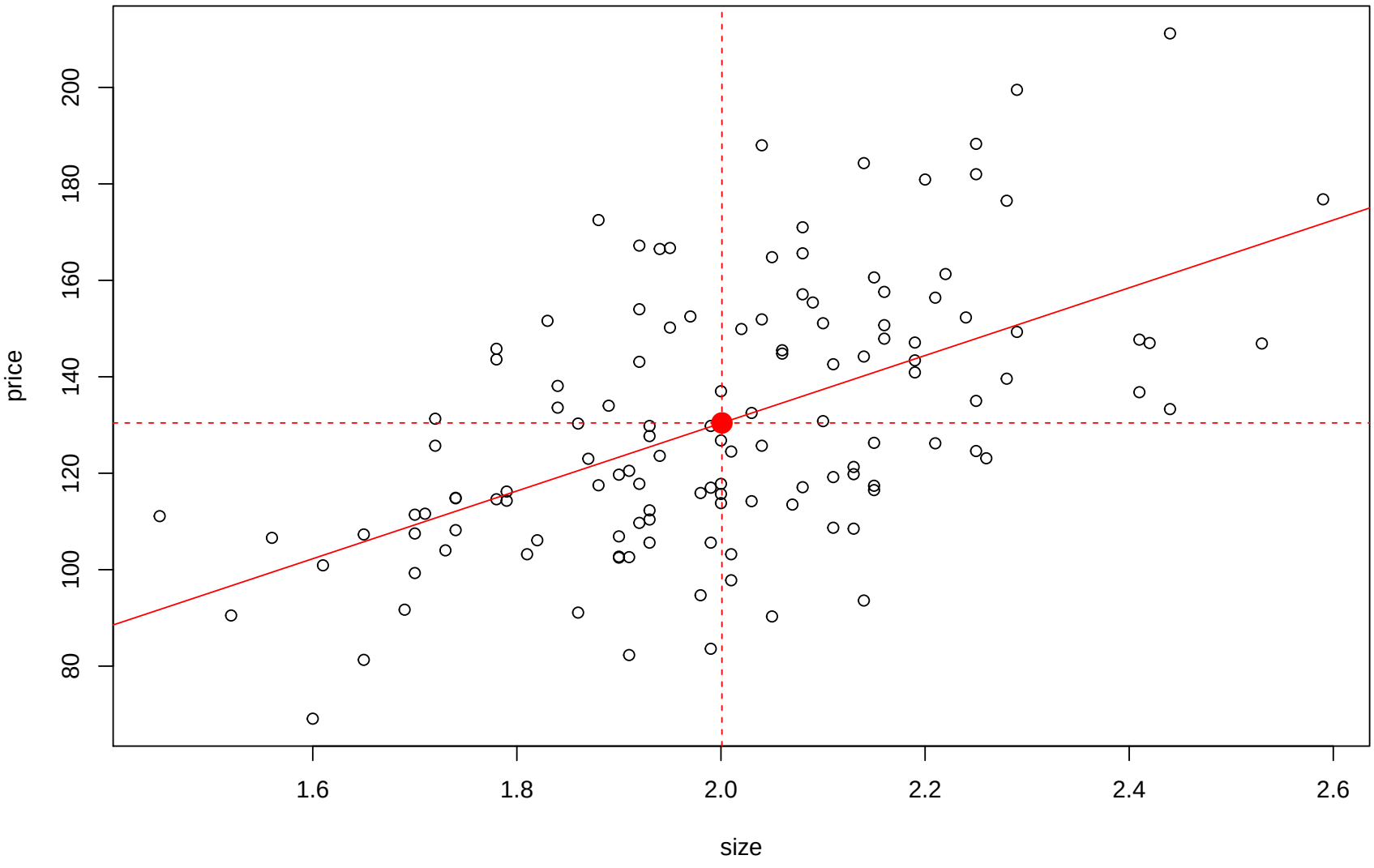
Least squares regression.

The values that give you the minimum are:

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$a = \bar{y} - b\,\bar{x}.$$

$a = \bar{y} - b\,\bar{x}$ is pretty intuitive if we just rearrange it to $\bar{y} = a + b\,\bar{x}$



There is not a simple story for b but we can usefully relate to the correlation. After all, both b and r are about the relationship between y and x.

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = r_{xy}\,\frac{s_y}{s_x}$$

Note that the correlation is symmetric: $r_{xy} = r_{yx}$

while regressing y on x does not give you the same slope as regressing x on y.

Summary stats and regression line for the Housing data:

Average size is 2.

Average price is 130.

cor(size,price) = 0.55298

sd size is 0.2116.

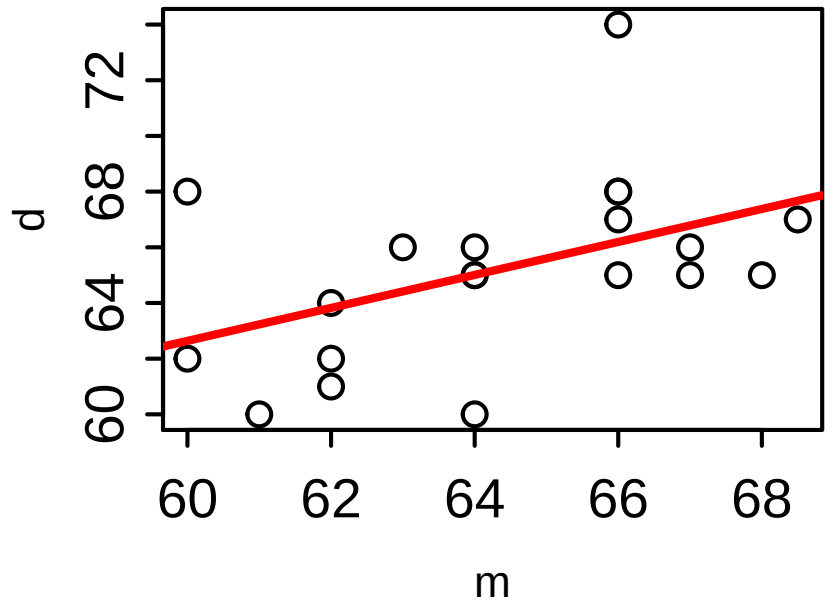
sd price is 26.8688.

b = 0.553*26.8688/0.2116
= 70.2195

a= 130-70.2195*2 = -10.439

Regression Towards the Mean

Let's look at the mother/daughter height data again.




```
##
## Call:
## lm(formula = d ~ m, data = hd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0075 -1.7939 -0.3231  1.1405  7.8080
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.1060    17.0333   1.591   0.1289
## m             0.5922     0.2646   2.238   0.0381 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.972 on 18 degrees of freedom
```

Now note that in general

$$\begin{aligned}\hat{y} &= a + bx_p \\ &= (\bar{y} - b\bar{x}) + bx_p \\ &= \bar{y} + b(x_p - \bar{x})\end{aligned}$$

So,

$$\hat{y} - \bar{y} = b\,(x_p - \bar{x})$$

This form relates how much y is predicted to be above the average of the y values in our data given how much x_p is above the average of x values in the data.

For our m/d data we have

$$\hat{d} - \bar{d} = .6\,(m_p - \bar{m})$$

If the mother is above average in height then we predict that the daughter is above average in height *but not by as much !!!*

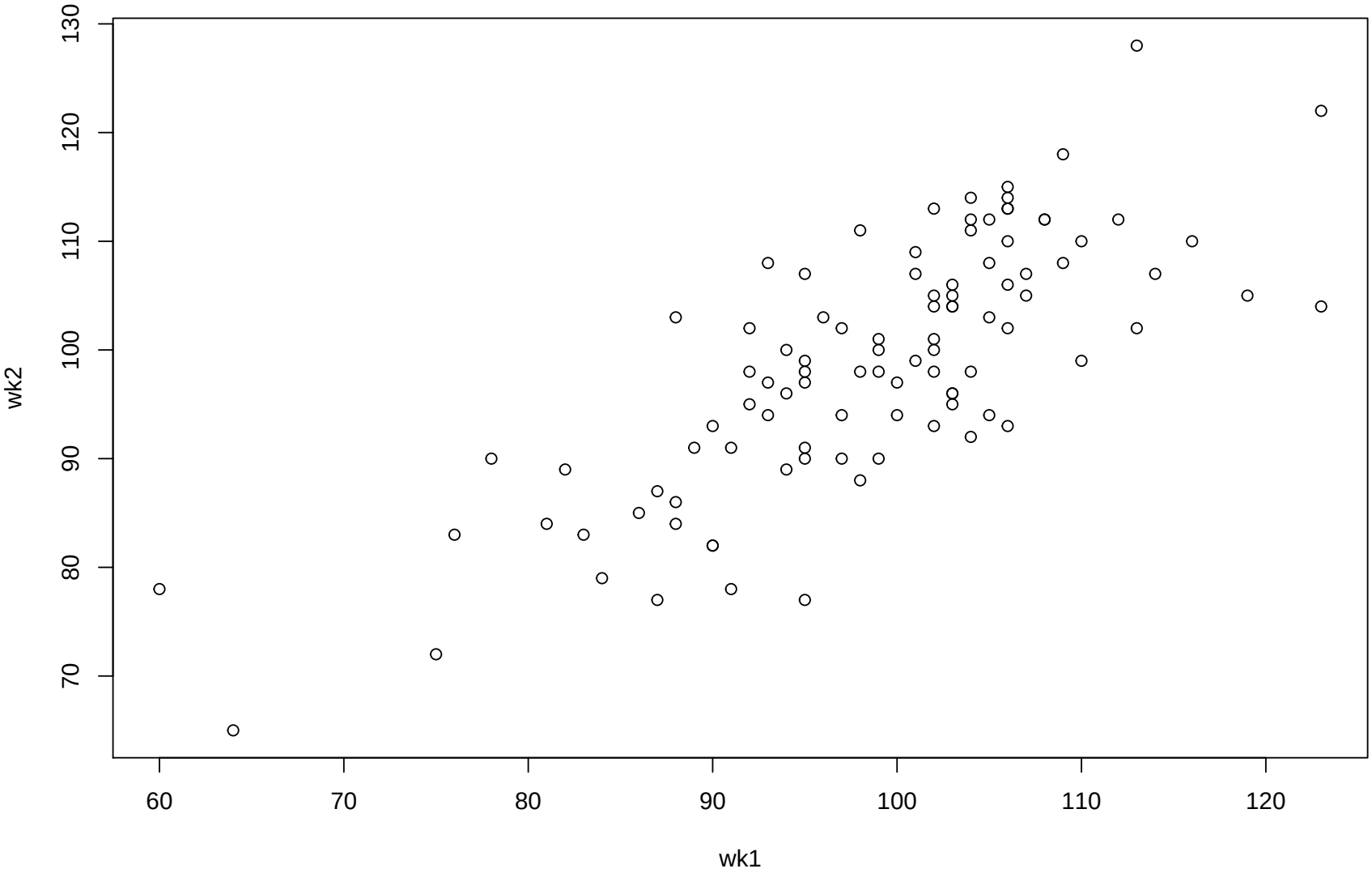
We have **regression towards the mean**.

Weekly performance data.

```
##   wk1 wk2
## 1  95  98
## 2 102 104
## 3  95  90
## 4 101 107
## 5  89  91
## 6 100  97
```

Each row corresponds to employee.

wk1 is performance in week 1.
wk2 is performance in week 2.


wk2
130
120
110
100
90
80
70
60
70
80
90
100
110
120
wk1

```
##
## Call:
## lm(formula = wk2 ~ wk1, data = pd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8013  -5.3820   0.7783   4.7599  17.3579
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.47514    6.78361   2.576   0.0115 *
## wk1          0.82449    0.06844  12.046   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.286 on 98 degrees of freedom
## Multiple R-squared:  0.5969, Adjusted R-squared:  0.5928
## F-statistic: 145.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

Suppose we have been leaving the employees alone but not institute a policy where we berate the below average performers.

What will we observe?

Questions

- how can we use the other variables (besides just size) to predict the price of a house?
- how sure are we about a prediction: what is the $\pm$?
- not every relationship is linear, what do I do in the nonlinear case ?
- what is the difference between correlation and causation anyway?
- what are all the other numbers of the regression output?
- where did Ken Griffey Jr. come from?