McCOMBS SCHOOL OF BUSINESS
Salem Center for Policy

Prediction

David Puelz

September 28, 2021

Outline

Simple linear regression

Multiple linear regression

Causal interpretation and extensions

Regression: General introduction

Regression analysis is the most widely used statistical tool for understanding relationships among variables

It provides a conceptually simple method for investigating functional relationships between one or more factors and an outcome of interest

The relationship is expressed in the form of an equation or a model connecting the response or dependent variable and one or more explanatory or predictor variable

Why?

Straight-up **prediction**:

- How much will I sell my house for?

Explanation and understanding:

- What is the impact of economic freedom on growth?

Example 1: Predicting house prices

Problem:

- Predict market price based on observed characteristics

Solution:

- Look at property sales data where we know the price and some observed characteristics.
- Build a decision rule that predicts price as a function of the observed characteristics.

Predicting house prices

Q: What characteristics do we use?

We have to define the variables of interest and develop a specific quantitative measure of these variables ...

Many factors or variables affect the price of a house:

- size
- number of baths
- garage, air conditioning, etc
- neighborhood

Predicting house prices

To keep things super simple, let's focus only on size. The value

that we seek to predict is called the dependent (or output) variable, and we denote this:

- Y = price of house (e.g. thousands of dollars)

The variable that we use to guide prediction is the explanatory (or input) variable, and this is labeled

- X = size of house (e.g. thousands of square feet)

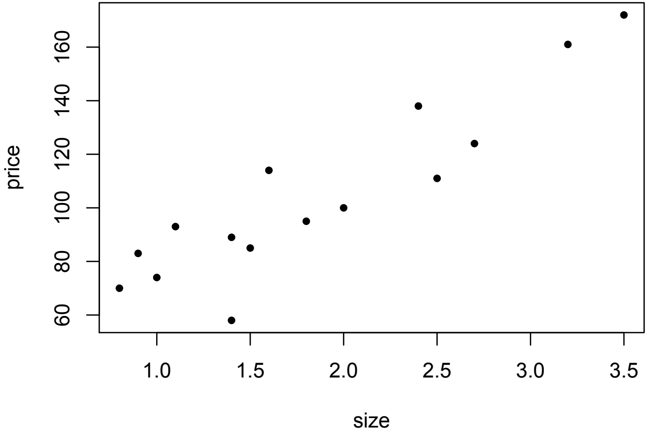
Predicting house prices

What does this data look like?

Size	Price
0.80	70
0.90	83
1.00	74
1.10	93
1.40	89
1.40	58
1.50	85
1.60	114
1.80	95
2.00	100
2.40	138
2.50	111
2.70	124
3.20	161
3.50	172

Predicting house prices

It is much more useful to look at a scatterplot



In other words, view the data as points in the $X \times Y$ plane.

Regression model

Y = response or outcome variable
X = explanatory or input variables

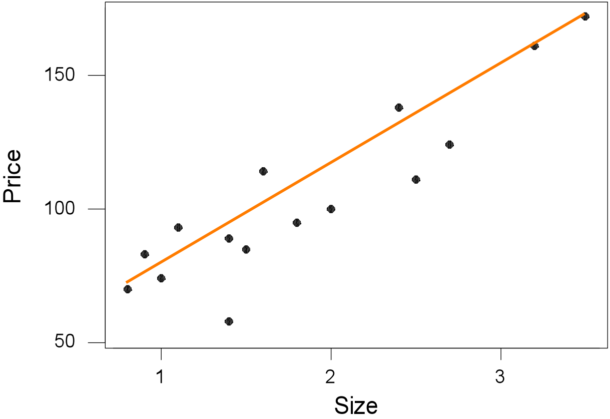
A linear relationship is written

$$Y = b_0 + b_1 X + e$$

Linear prediction

There seems to be a linear relationship between price and size:

As size goes up, price goes up.



Linear prediction

Recall that the equation of a line is:

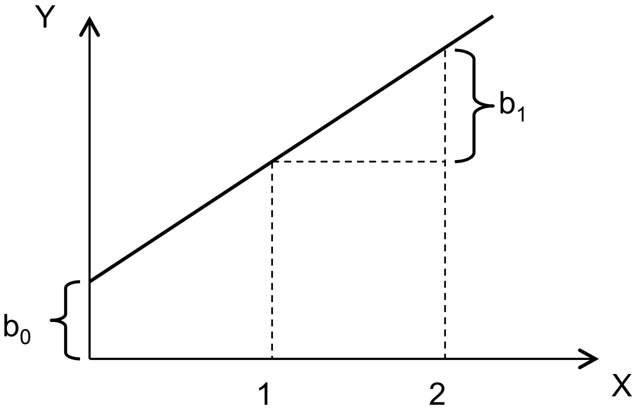
$$Y = b_0 + b_1 X$$

Where b_0 is the **intercept** and b_1 is the **slope**.

→ The **intercept** value is in units of Y ($1,000)

→ The **slope** is in units of Y *per* units of X ($1,000/1,000 sq ft)

Linear prediction



$Y = b_0 + b_1X$

Linear prediction

Q: How to find the “best line”?

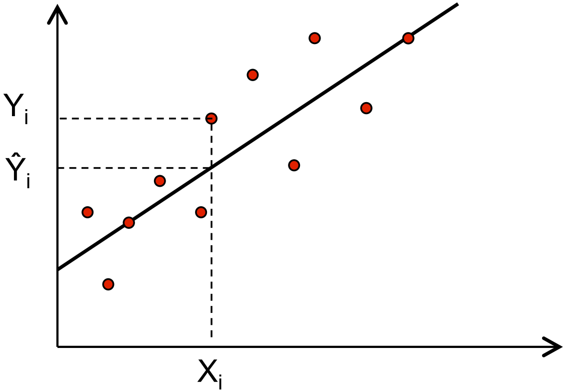
We desire a strategy for estimating the slope and intercept parameters in the model $\hat{Y} = b_0 + b_1 X$

A reasonable way to fit a line is to minimize the amount by which the fitted value differs from the actual value.

This amount is called the residual.

Linear prediction

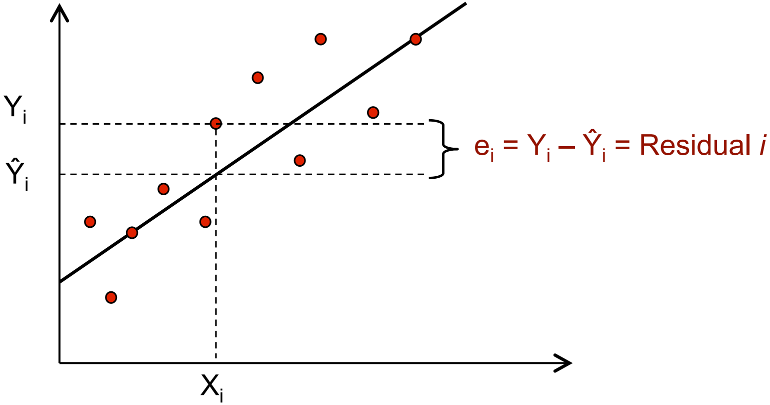
What is the “fitted value”?



The dots are the observed values and the line represents our fitted values given by $\hat{Y}_i = b_0 + b_1 X_1$.

Linear prediction

What is the “residual”’ for the *i*th observation?



We can write $Y_i = \hat{Y}_i + (Y_i - \hat{Y}_i) = \hat{Y}_i + e_i$.

Least squares

Ideally, we want to minimize the size of all residuals:

- If they were all zero we would have a perfect line.
- Trade-off between moving closer to some points and at the same time moving away from other points.

Least squares

Ideally, we want to minimize the size of all residuals:

- If they were all zero we would have a perfect line.
- Trade-off between moving closer to some points and at the same time moving away from other points.

The line fitting process:

- Give weights to all of the residuals.
- Minimize the “total” of residuals to get best fit.

Least squares

Ideally, we want to minimize the size of all residuals:

- If they were all zero we would have a perfect line.
- Trade-off between moving closer to some points and at the same time moving away from other points.

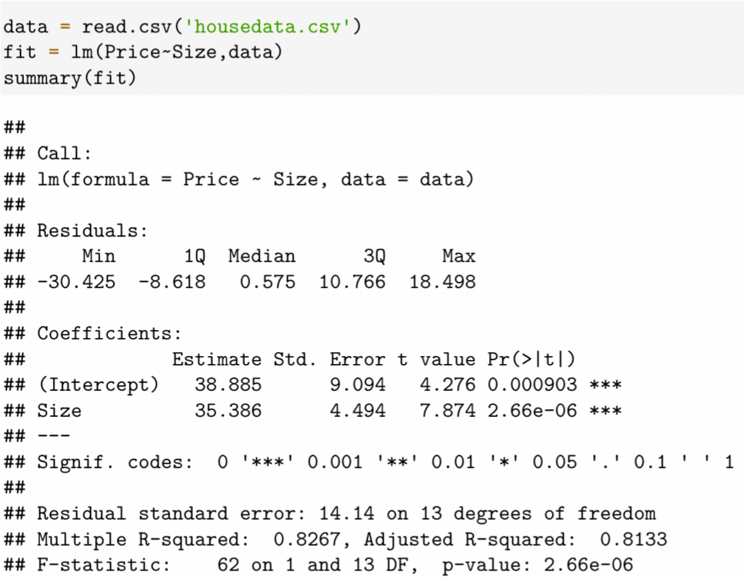
The line fitting process:

- Give weights to all of the residuals.
- Minimize the "total" of residuals to get best fit.

  Least Squares chooses b_0 and b_1 to minimize $\sum_{i=1}^{N} e_i^2$

$$\sum_{i=1}^{N} e_i^2 = e_1^2 + e_2^2 + \cdots + e_N^2 = (Y_1 - \hat{Y}_1)^2 + (Y_2 - \hat{Y}_2)^2 + \cdots + (Y_N - \hat{Y}_N)^2$$

Least squares – R output



```
data = read.csv('housedata.csv')
fit = lm(Price~Size,data)
summary(fit)
```

```
##
## Call:
## lm(formula = Price ~ Size, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.425  -8.618   0.575  10.766  18.498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.885      9.094   4.276 0.000903 ***
## Size          35.386      4.494   7.874 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.14 on 13 degrees of freedom
## Multiple R-squared:  0.8267, Adjusted R-squared:  0.8133
## F-statistic:    62 on 1 and 13 DF,  p-value: 2.66e-06
```

Example 2: Offensive performance in baseball

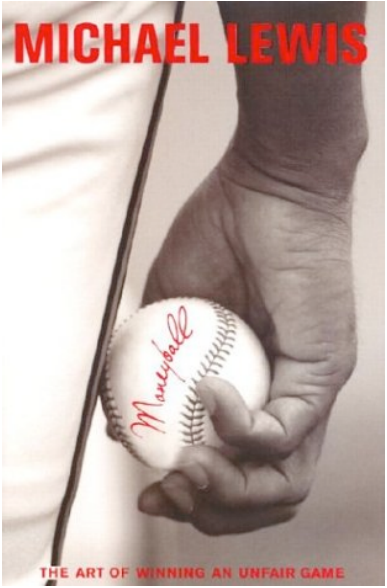
Problems:

- Evaluate/compare traditional measures of offensive performance
- Help evaluate the worth of a player

Solutions:

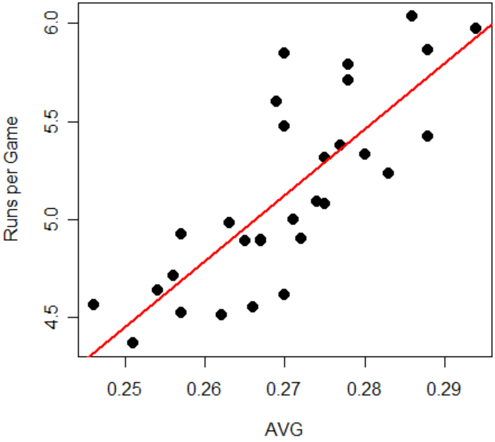
- Compare *prediction rules* that forecast runs as a function of either AVG (batting average), SLG (slugging percentage – total bases divided by at bats) or OBP (on base percentage)

Example 2: Offensive performance in baseball

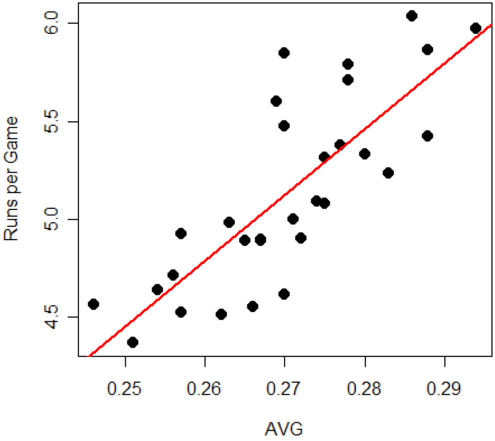


Baseball data – using AVG

Each observation corresponds to a team in MLB. Each quantity is the average over a season.



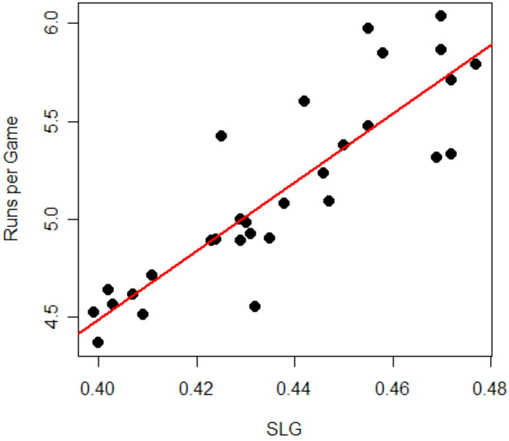
Baseball data – using AVG



Y = runs per game; X = AVG (average)

LS fit: Runs/Game = -3.93 + 33.57 AVG

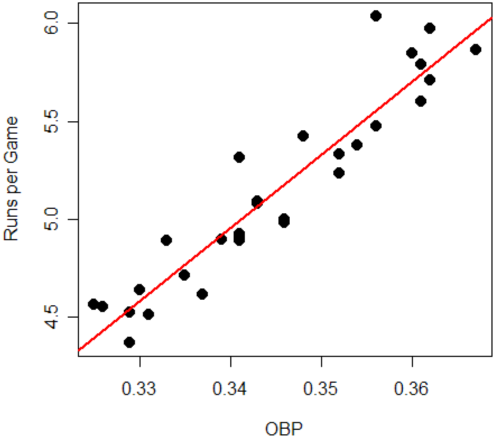
Baseball Data – using SLG



Y = runs per game; X = SLG (slugging percentage)

LS fit: Runs/Game = -2.52 + 17.54 SLG

Baseball Data – using OBP



Y = runs per game; X = OBP (on base percentage)

LS fit: Runs/Game = -7.78 + 37.46 OBP

Baseball data

- What is the best prediction rule?

- Let's compare the predictive ability of each model using the average squared error

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2} = \left(\frac{\sum_{i=1}^{N} \left(\widehat{\text{Runs}_i} - \text{Runs}_i \right)^2}{N} \right)^{\frac{1}{2}}$$

Place your money on OBP!!!

	Root Mean Squared Error
AVG	0.29
SLG	0.23
OBP	0.16

More on least squares

Remember how we get the slope (b_1) and intercept (b_0). We minimize the sum of squared prediction errors.

The formulas for b_0 and b_1 that minimize the least squares criterion are:

$$b_1 = r_{xy} \times \frac{s_y}{s_x} \qquad b_0 = \bar{Y} - b_1 \bar{X}$$

where,

- $\bar{X}$ and $\bar{Y}$ are the sample mean of X and Y
- $\mathrm{corr}(x, y) = r_{xy}$ is the sample correlation
- s_x and s_y are the sample standard deviation of X and Y

What are these numbers in the formula?

- Sample Mean: measure of centrality

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

- Sample Variance: measure of spread

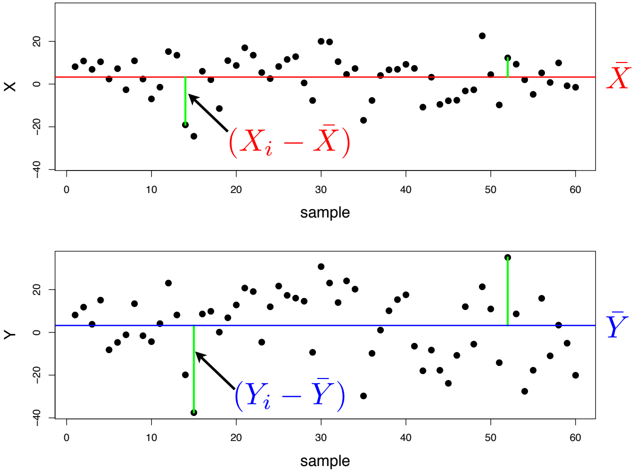
$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$$

- Sample Standard Deviation:

$$s_y = \sqrt{s_y^2}$$

Visual: standard deviation

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$$

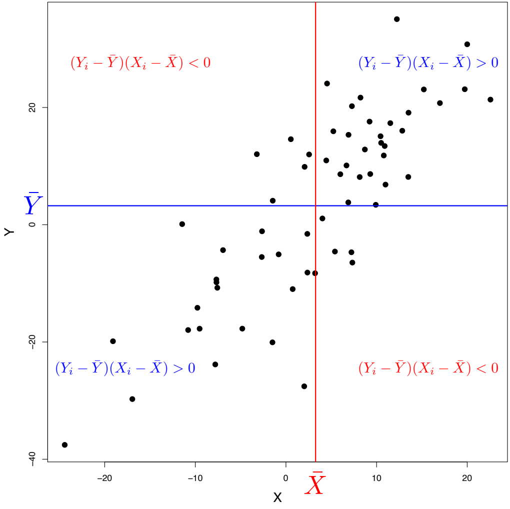


$s_x = 9.7$ $s_y = 15.98$

Visual: Covariance

Measure the **direction** and **strength** of the linear relationship between Y and X

$$\text{cov}(Y,X) = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{n-1}$$



- $s_y = 15.98,\ s_x = 9.7$
- $\text{cov}(X,Y) = 125.9$

How do we interpret that?

A standardized measure: Correlation

Correlation is the standardized covariance:

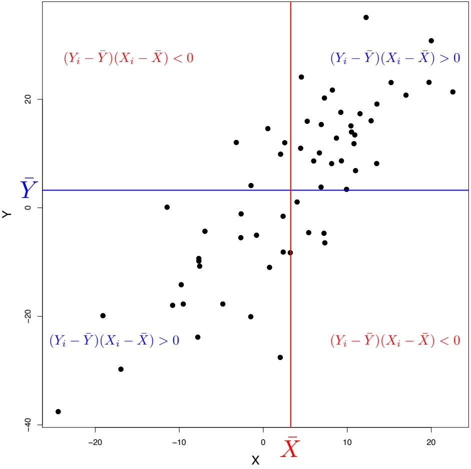
$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{s_x^2 s_y^2}} = \frac{\mathrm{cov}(X, Y)}{s_x s_y}$$

The correlation is scale invariant and the units of measurement don't matter: It is always true that $-1 \leq \mathrm{corr}(X, Y) \leq 1$.

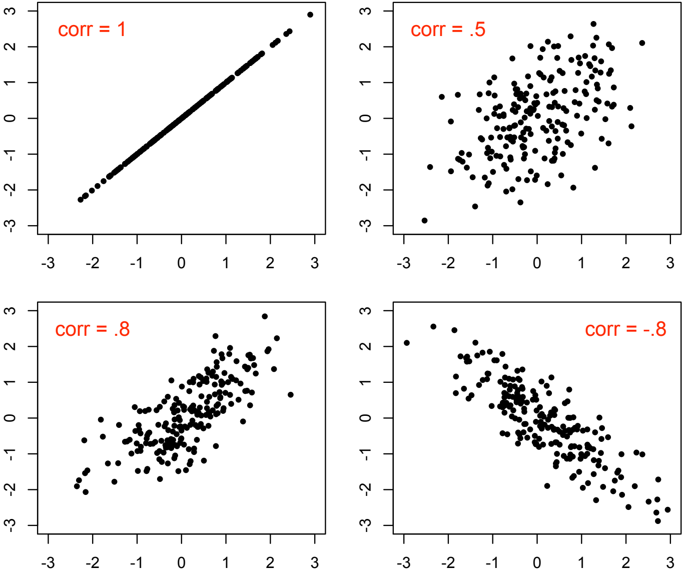
This gives the direction (negative or positive) and strength $(0 \rightarrow 1)$ of the linear relationship between X and Y.

Correlation

$$\text{corr}(Y, X) = \frac{\text{cov}(X, Y)}{\sqrt{s_x^2 s_y^2}} = \frac{\text{cov}(X, Y)}{s_x s_y} = \frac{125.9}{15.98 \times 9.7} = 0.812$$



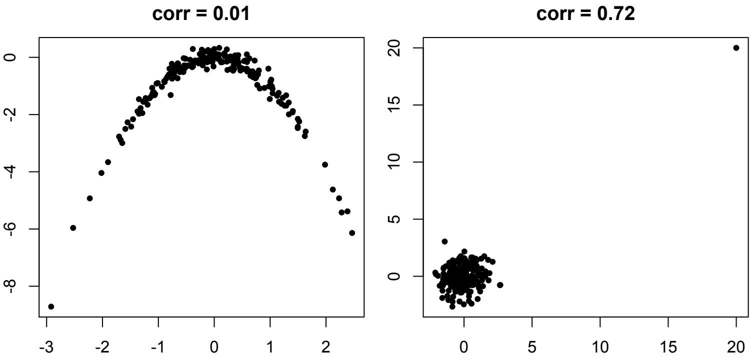
Correlation



Correlation

Only measures linear relationships:

$\text{corr}(X, Y) = 0$ does not mean the variables are not related!



Also be careful with influential observations. Check out `cor()` in `R`.

Back to least squares

Intercept:

$$b_0 = \bar{Y} - b_1\bar{X} \Rightarrow \bar{Y} = b_0 + b_1\bar{X}$$

The point $(\bar{X}, \bar{Y})$ is on the regression line!

Least squares finds the point of means and rotates the line through that point until getting the "right" slope

Slope:

$$\begin{aligned} b_1 = \text{corr}(X, Y) \times \frac{s_Y}{s_X} &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ &= \frac{\text{cov}(X, Y)}{\text{var}(X)} \end{aligned}$$

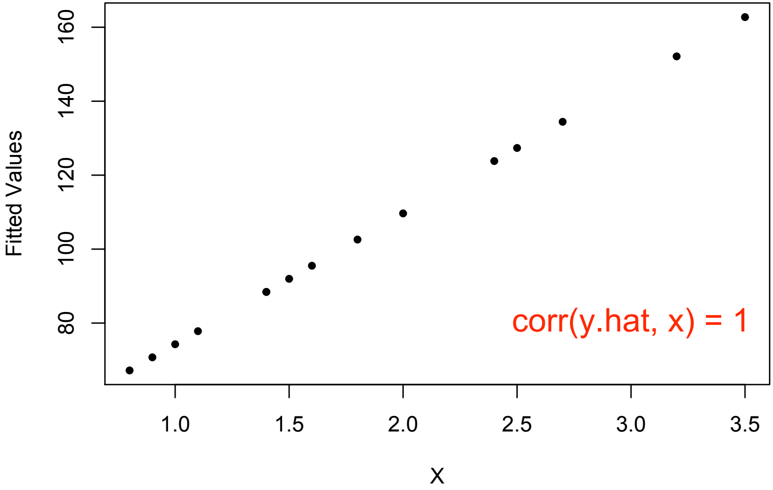
So, the right slope is the correlation coefficient times a scaling factor that ensures the proper units for b_1

More on least squares

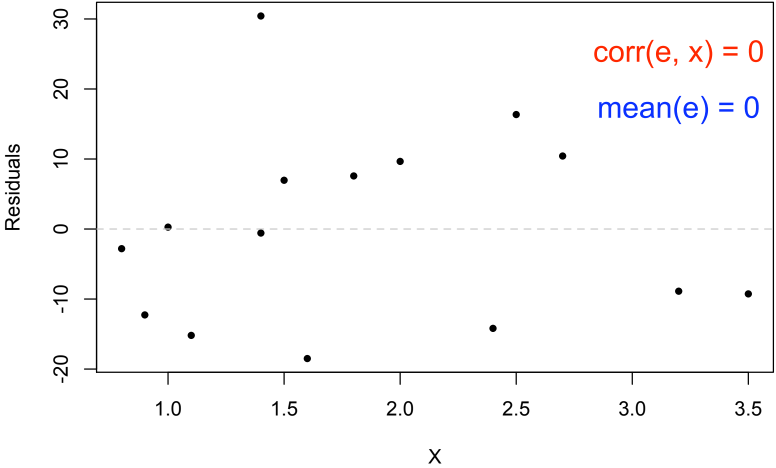
From now on, terms “fitted values” ($\hat{Y}_i$) and “residuals” (e_i) refer to those obtained from the least squares line.

The fitted values and residuals have some special properties. Lets look at the housing data analysis to figure out what these properties are...

The fitted values and X

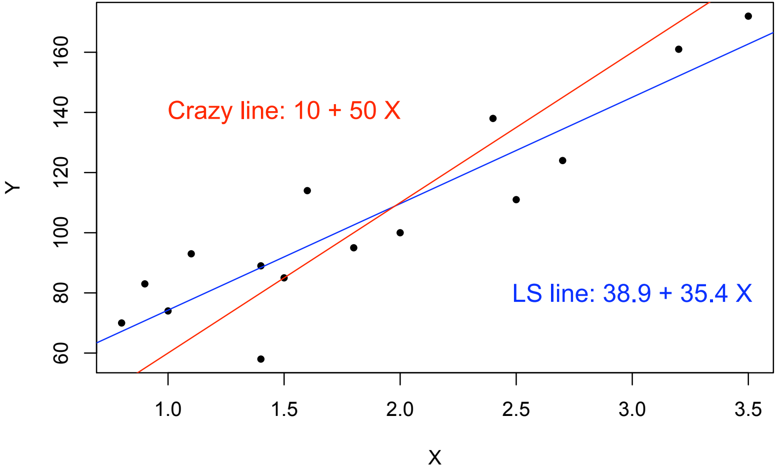


The residuals and X



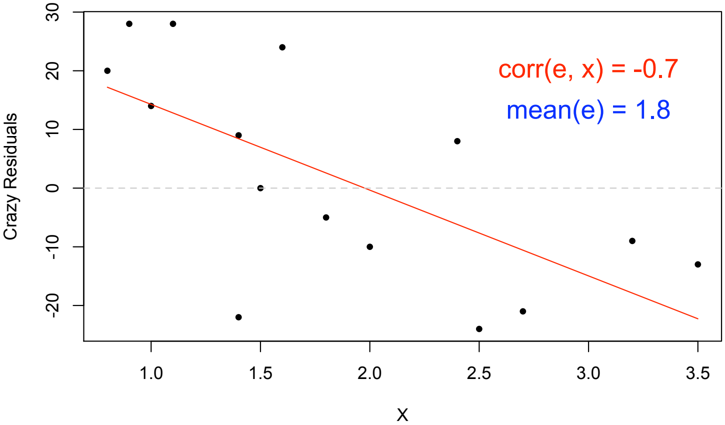
Why?

What is the intuition for the relationship between $\hat{Y}$ and e and X? Lets consider some "crazy"alternative line:



Fitted values and residuals

This is a bad fit! We are underestimating the value of small houses and overestimating the value of big houses.



Clearly, we have left some predictive ability on the table!

Fitted Values and Residuals

As long as the correlation between e and X is non-zero, we could always adjust our prediction rule to do better.

We need to exploit all of the predictive power in the X values and put this into $\hat{Y}$, leaving no "*Xness*" in the residuals.

In summary: $Y = \hat{Y} + e$ where:

- $\hat{Y}$ is "made from X"; corr$(X, \hat{Y}) = 1$.
- e is unrelated to X; corr$(X, e) = 0$.