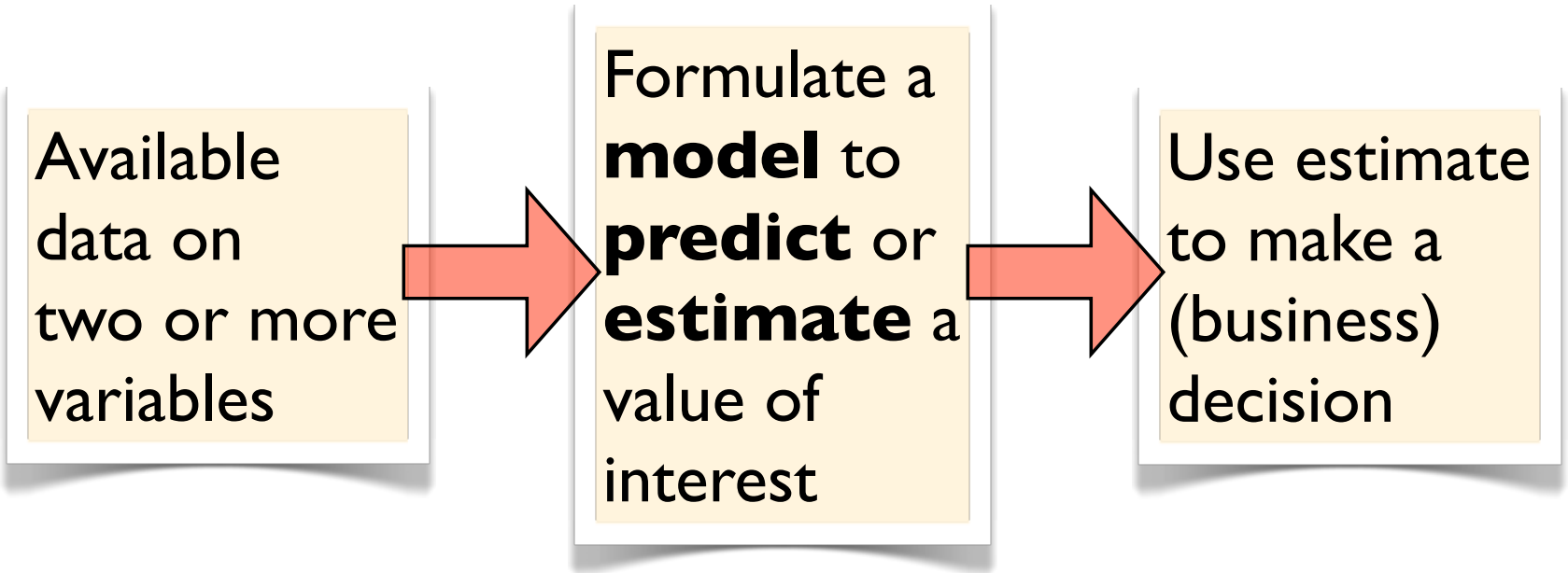
Business Statistics 41000

Simple Linear Regression

Mladen Kolar

The basic problem



Regression: What is it?

Simply: The most widely used statistical tool for understanding relationships among variables

A conceptually simple method for investigating 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 outcome to the factors

Regression in business

Optimal portfolio choice:

- Predict the future joint distribution of asset returns
- Construct an optimal portfolio (choose weights)

Determining price and marketing strategy:

- Estimate the effect of price and advertisement on sales
- Decide what is optimal price and ad campaign

Credit scoring model:

- Predict the future probability of default using known characteristics of borrower
- Decide whether or not to lend (and if so, how much)

Regression in everything

Straight prediction questions:

- What price should I charge for my car?
- What will the interest rates be next month?
- Will this person like that movie?

Explanation and understanding:

- Does your income increase if you get an MBA?
- Will tax incentives change purchasing behavior?
- Is my advertising campaign working?

Example: pickup truck prices on Craigslist

We have 4 dimensions to consider.

```
data <- read.csv("pickup.csv")
names(data)
```

```
## [1] "year"  "miles" "price" "make"
```

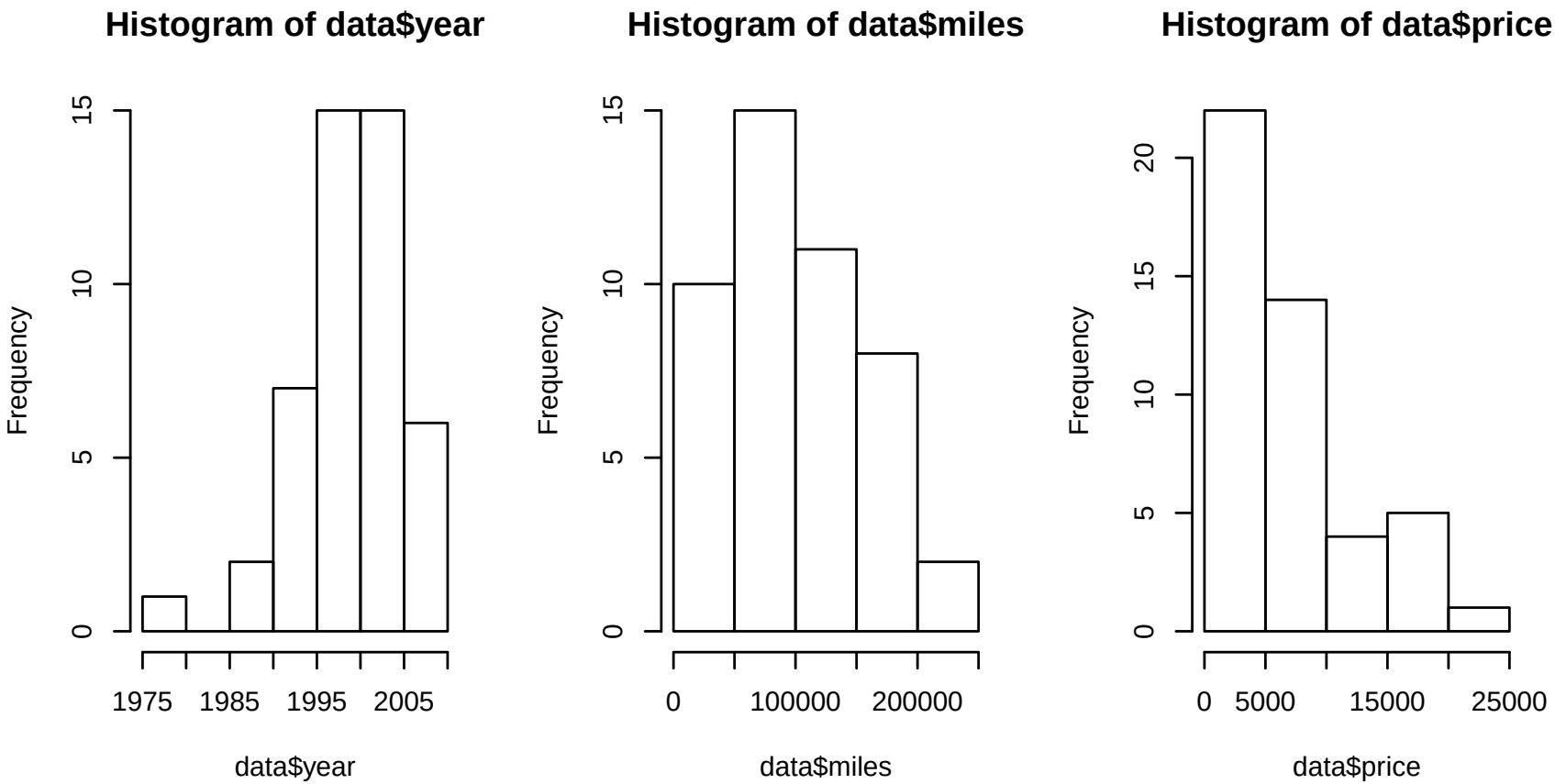
A simple summary is

```
summary(data)
```

```
##       year          miles            price          make
##  Min.   :1978   Min.   :  1500   Min.   : 1200   Dodge:10
##  1st Qu.:1996   1st Qu.: 70958   1st Qu.: 4099   Ford :12
##  Median :2000   Median : 96800   Median : 5625   GMC  :24
##  Mean   :1999   Mean   :101233   Mean   : 7910
##  3rd Qu.:2003   3rd Qu.:130375   3rd Qu.: 9725
##  Max.   :2008   Max.   :215000   Max.   :23950
```

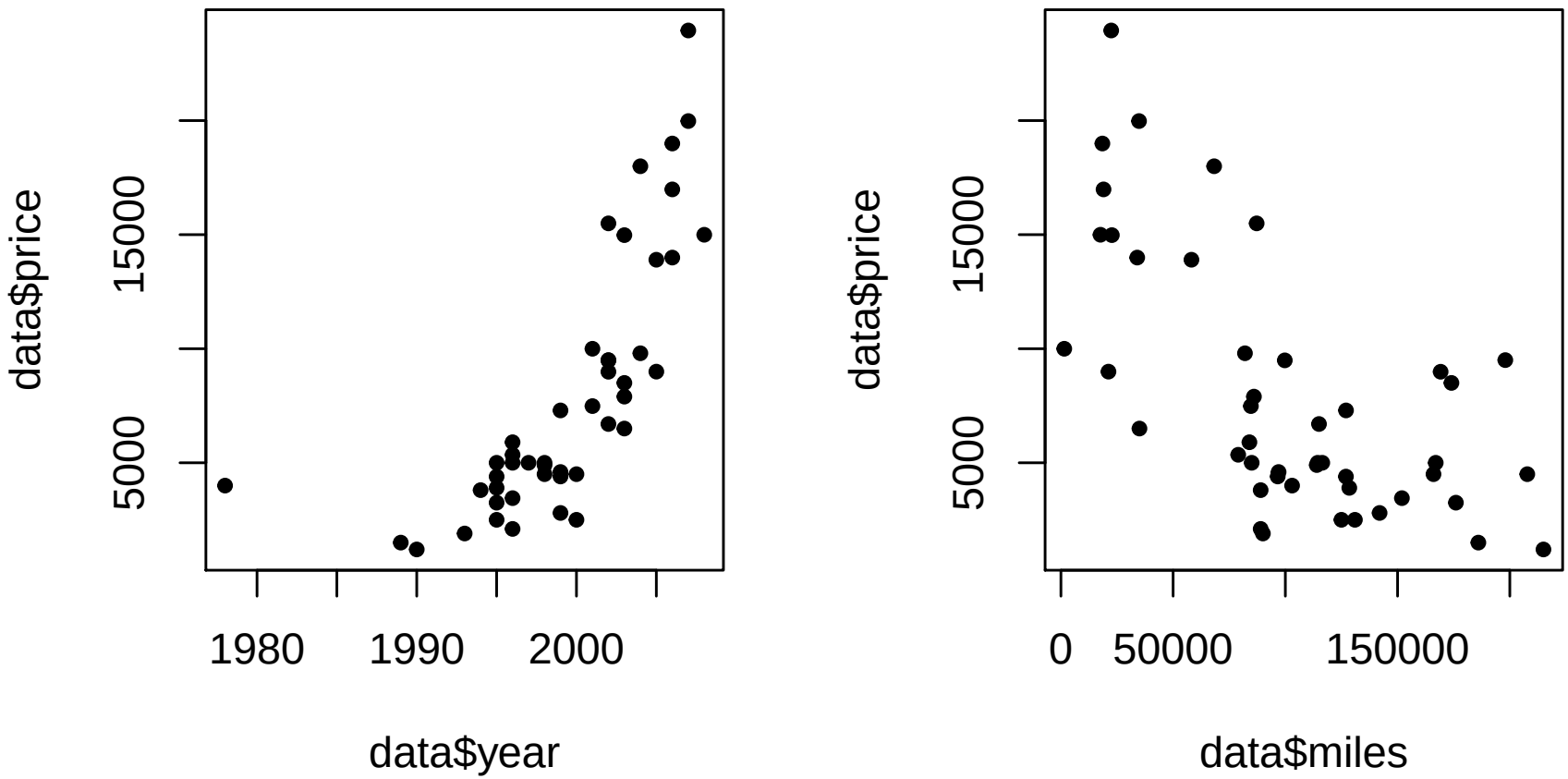
First, the simple histogram (for each continuous variable).

```
par(mfrow=c(1,3)) # break the plot into a 1x3 matrix
hist(data$year)
hist(data$miles)
hist(data$price)
```



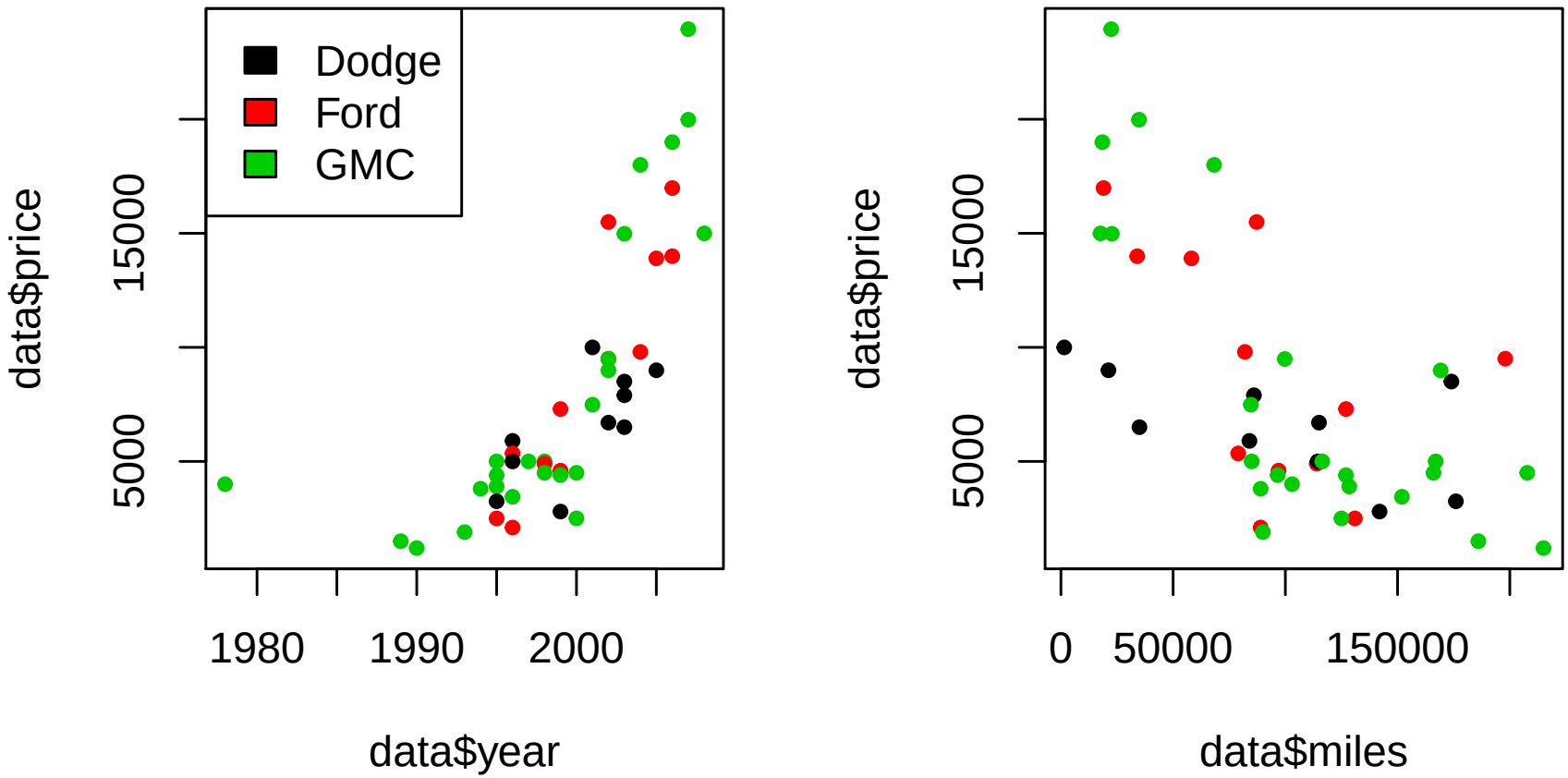
Let us use scatterplots to compare two dimensions.

```
par(mfrow=c(1,2))
plot(data$year, data$price, pch=20)
plot(data$miles, data$price, pch=20)
```



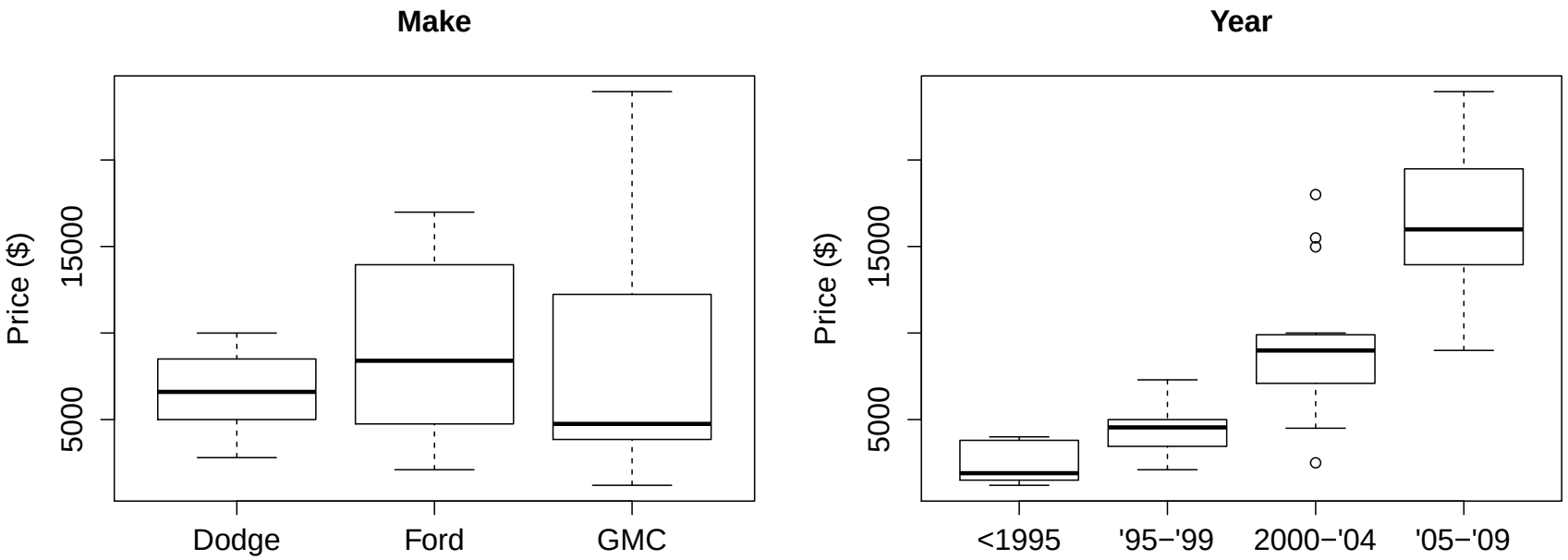
Add color to see another dimension.

```
par(mfrow=c(1,2))
plot(data$year, data$price, pch=20, col=data$make)
legend("topleft", levels(data$make), fill=1:3)
plot(data$miles, data$price, pch=20, col=data$make)
```



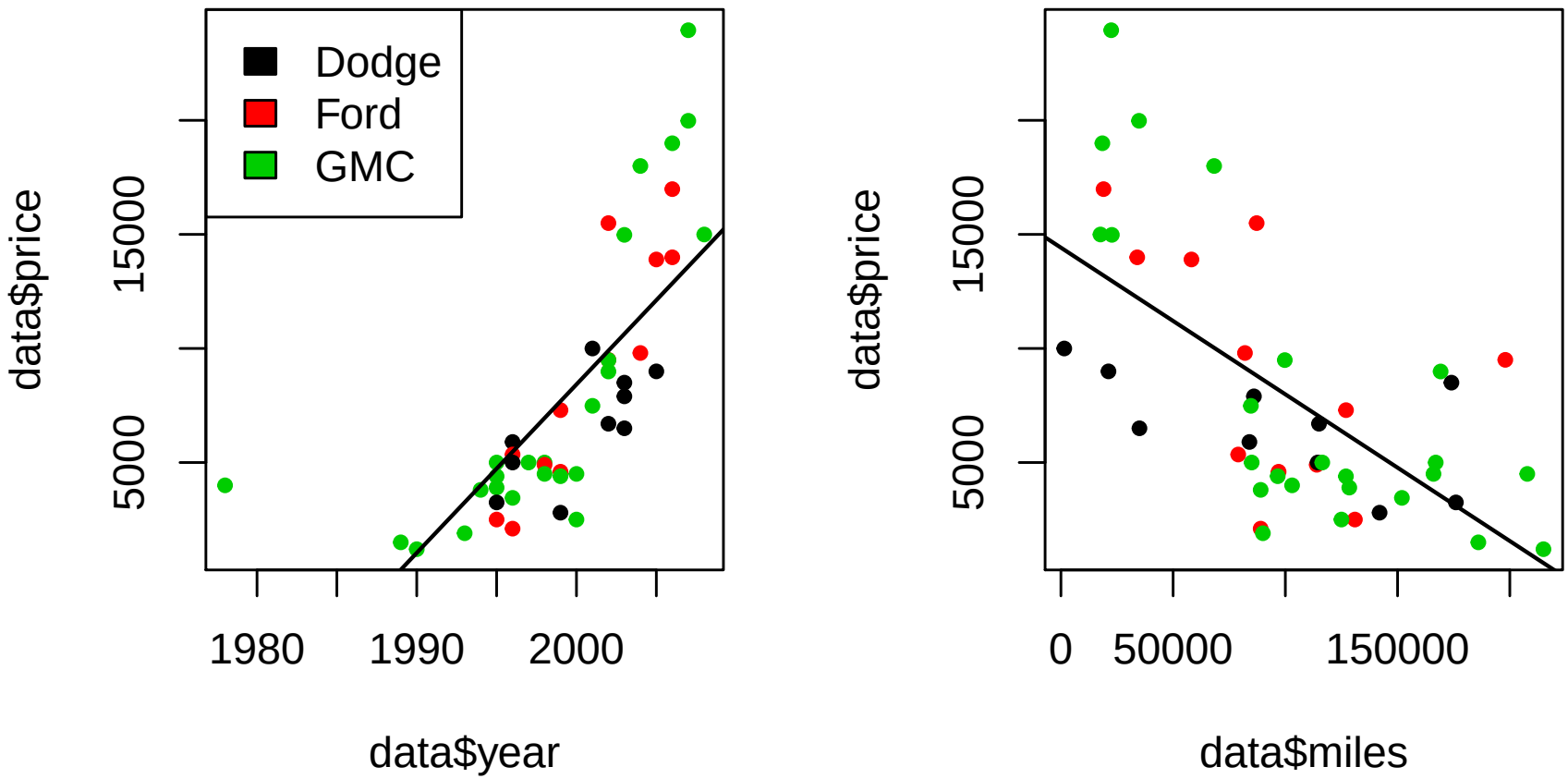
Boxplots are also super useful.

```
attach(data)
year_boxplot <- factor(1*(year<1995) + 2*(1995<=year & year<2000)
    + 3*(2000<=year & year<2005) + 4*(2005<=year & year<2009),
    labels=c("<1995", "'95-'99", "2000-'04", "'05-'09"))
par(mfrow=c(1,2))
boxplot(price ~ make, ylab="Price ($)", main="Make",
        cex.main=1.3, cex.lab=1.3, cex.axis=1.3)
boxplot(price ~ year_boxplot, ylab="Price ($)", main="Year",
        cex.main=1.3, cex.lab=1.3, cex.axis=1.3)
```



Regression is what we're really here for.

```
par(mfrow=c(1,2))
plot(data$year, data$price, pch=20, col=data$make)
abline(lm(price ~ year),lwd=1.5)
legend("topleft", levels(data$make), fill=1:3)
plot(data$miles, data$price, pch=20, col=data$make)
abline(lm(price ~ miles),lwd=1.5)
```



Conditional distributions

Regression models are really all about modeling the conditional distribution of Y given X.

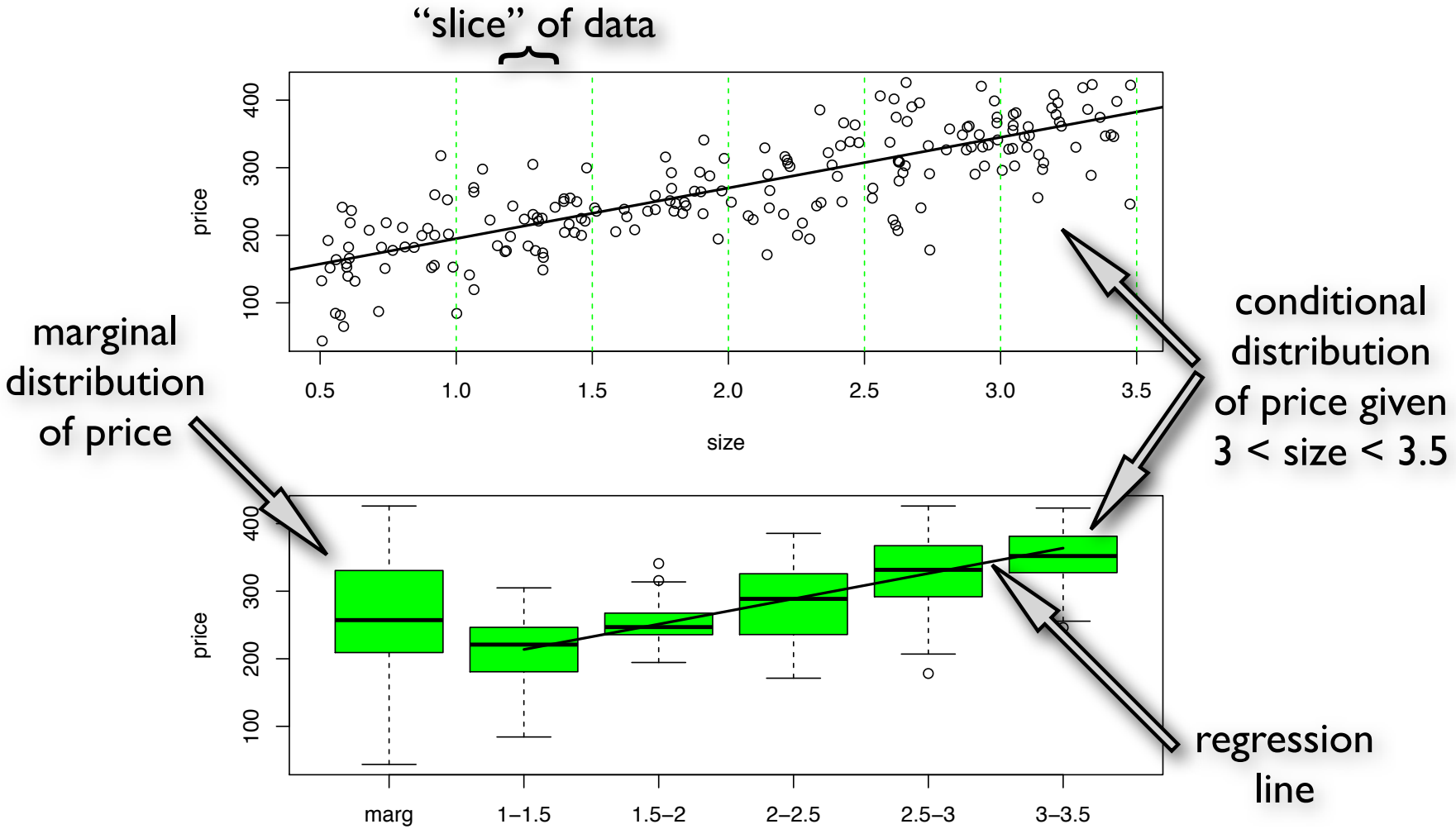
Why are conditional distributions important?

We want to develop models for forecasting. What we are doing is exploiting the information in the conditional distribution of Y given X.

The conditional distribution is obtained by **slicing** the point cloud in the scatterplot to obtain the distribution of Y conditional on various ranges of X values.

Conditional vs. marginal distribution

Consider a regression of house **price** on **size**:



Key observations from these plots:

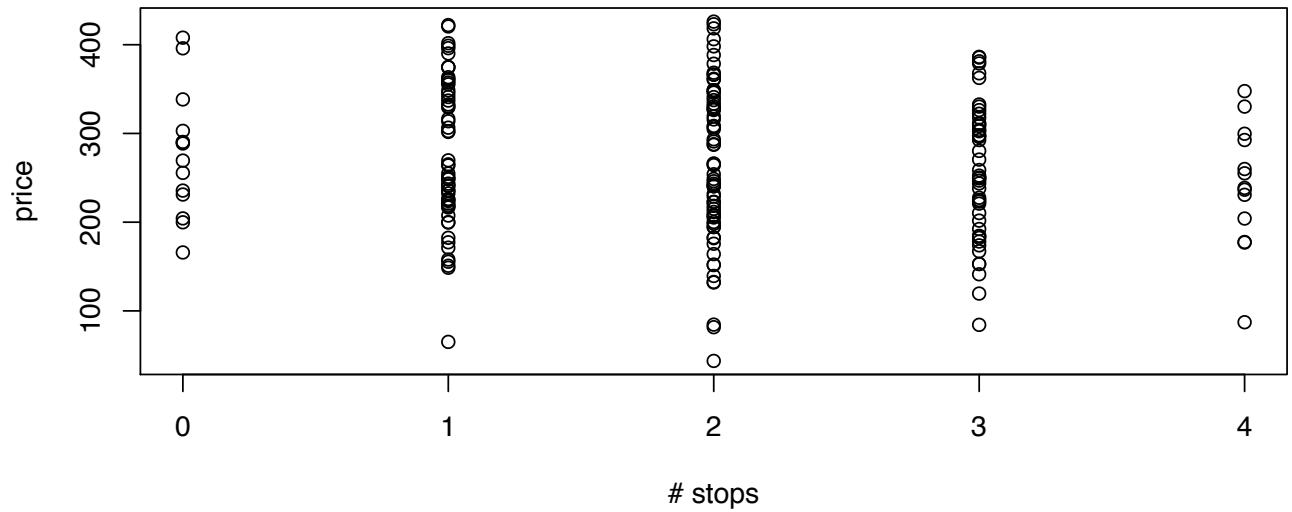
- Conditional distributions answer the forecasting problem: if I know that a house is between 1 and 1.5 1000 sq.ft., then the conditional distribution (second boxplot) gives me a point forecast (the mean) and prediction interval.
- The conditional means seem to line up along the regression line.
- The conditional distributions have much smaller dispersion than the marginal distribution.

This suggests two general points:

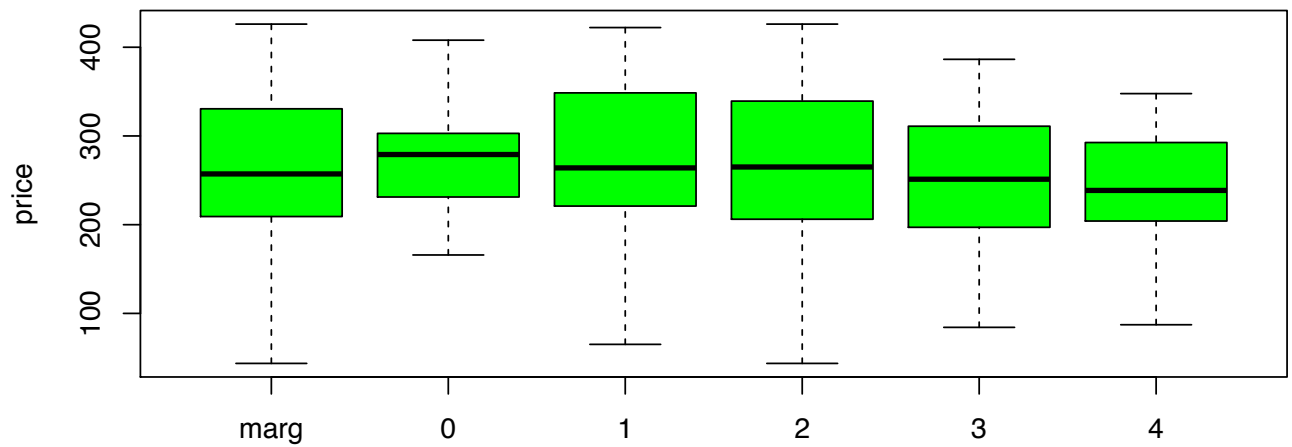
- If X has no forecasting power, then the marginal and conditionals will be the same.
- If X has some forecasting information, then conditional means will be different than the marginal or overall mean and the conditional standard deviation of Y given X will be less than the marginal standard deviation of Y.

Intuition from an example where X has no predictive power.

House price v. number of stop signs (Y) within a two-block radius of a house (X)



See that in this case the marginal and conditionals are not all that different



Linear regression model

Y = response or outcome variable
$X_1, X_2, X_3, \ldots, X_p$ = explanatory or input variables

A linear relationship is written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Here ε represents anything left over, not described by the line.

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.

$\Longrightarrow$ We have to define the variables of interest and develop a specific quantitative measure of these variables

What characteristics do we use?

- Many factors or variables affect the price of a house
  - size of house
  - number of baths
  - garage, air conditioning, etc.
  - size of land
  - location
- Easy to quantify price and size but what about other variables such as location, aesthetics, workmanship, etc?

To keep things super simple, let's focus only on `size` of the house.

The value that we seek to predict is called the **dependent (or output)** variable, and we denote this as

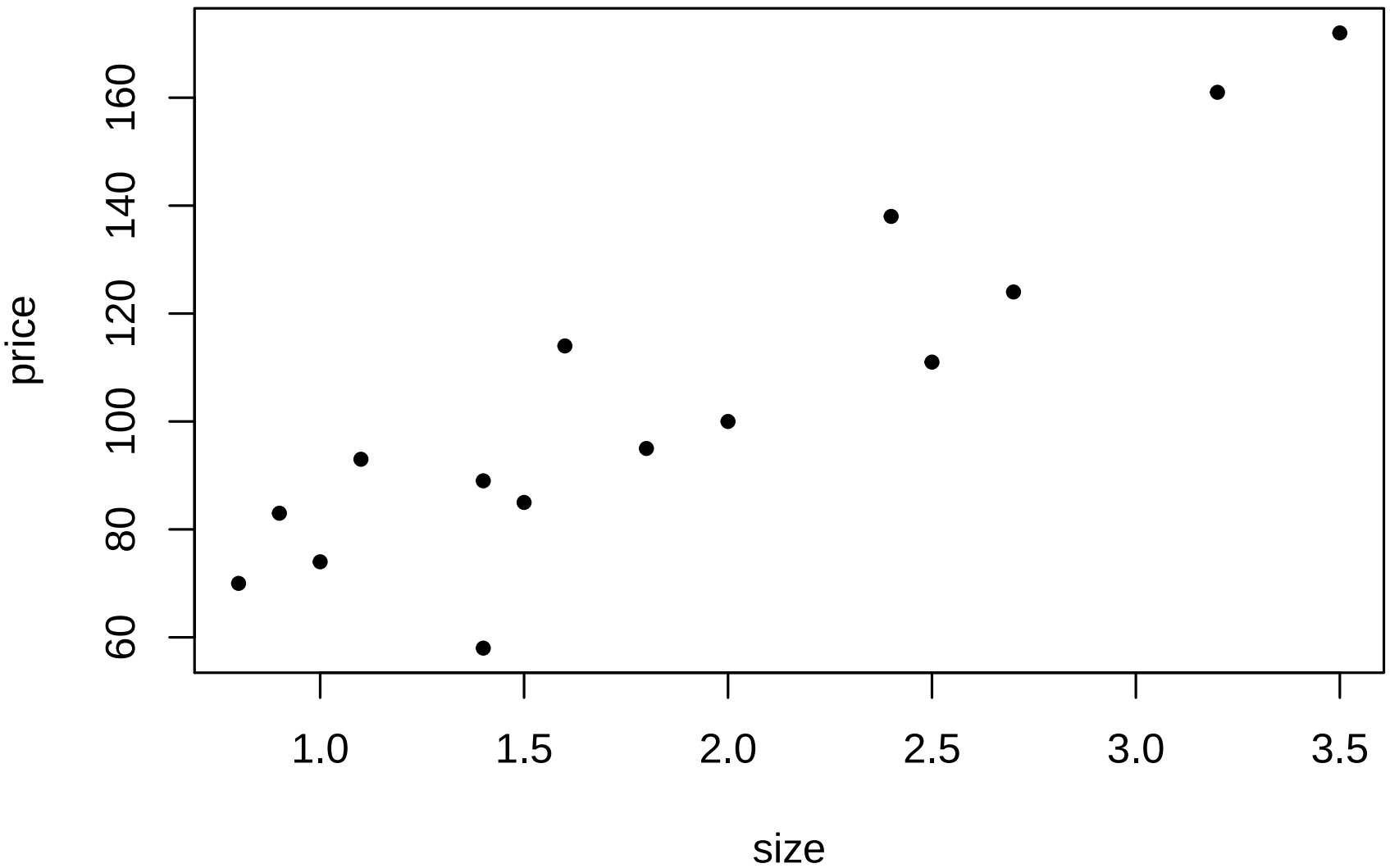
- Y = price of house (e.g. thousands of dollars)

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

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

What do the data look like?

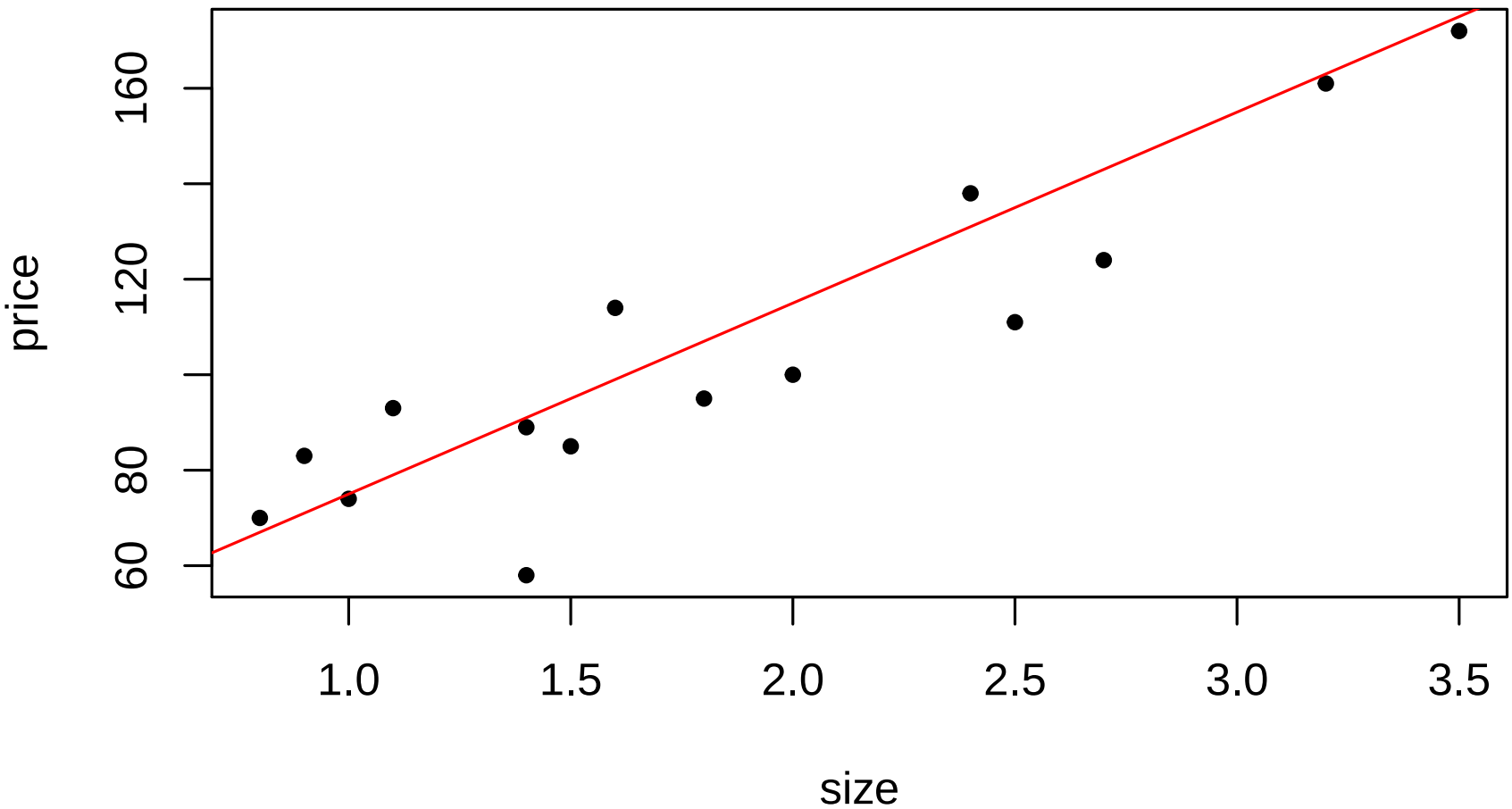
```
plot(size, price, pch=20)
```



Appears to be a linear relationship between `price` and `size`:

- as `size` goes up, `price` goes up.

```
plot(size, price, pch=20)
abline(35, 40, col="red")
```



The line shown was fit by the **eyeball** method.

Recall that the equation of a line is:

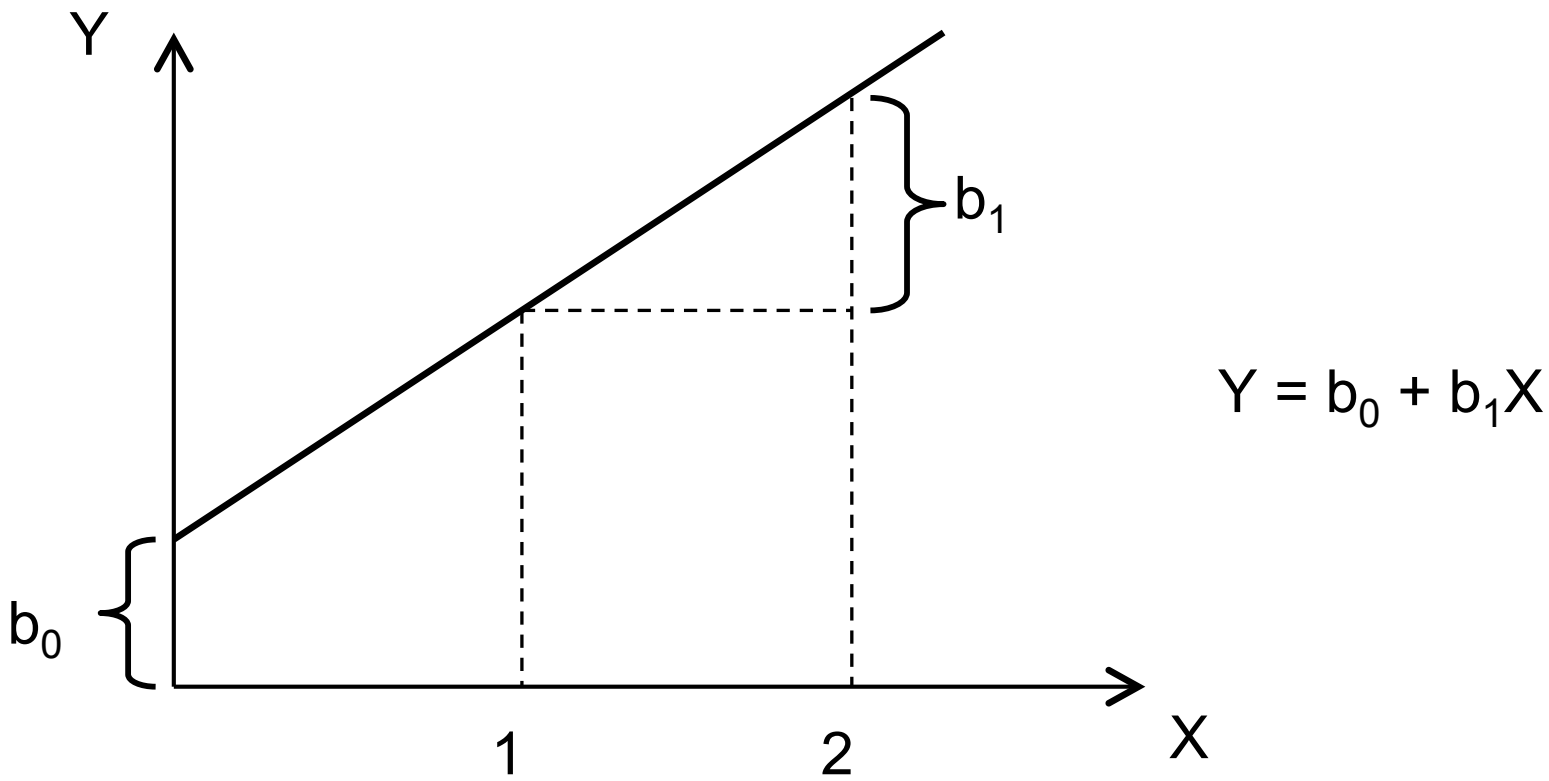
$$\hat{Y} = b_0 + b_1 X$$

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

In the house price example

- our *eyeball* line had $b_0 = 35$, $b_1 = 40$
- **predict** the `price` of a house when we know only `size`
  - just read the value off the line that we've drawn.
- 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).

Recall how the slope (b_1) and intercept (b_0) work together **graphically**



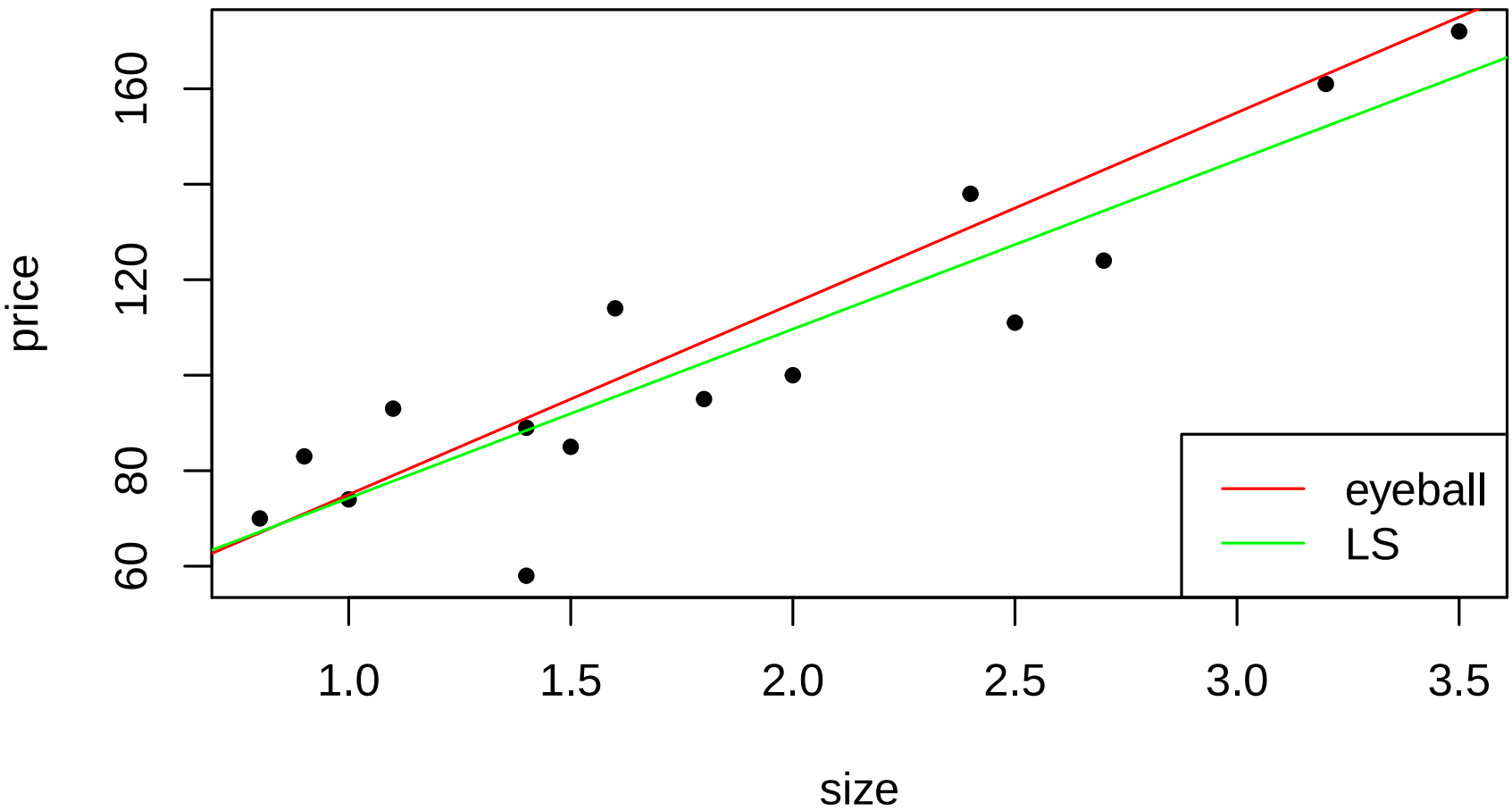
R's `lm` command provides a least squares fit.

```
(reg = lm(price ~ size))
```

```
##
## Call:
## lm(formula = price ~ size)
##
## Coefficients:
## (Intercept)         size
##        38.9         35.4
```

`lm` stands for **l**inear **m**odel; it'll be our workhorse

```
plot(size, price, pch=20)
abline(35, 40, col="red")
abline(reg, col="green")
legend("bottomright", c("eyeball", "LS"), col=c("red", "green"), lty=1)
```



The least squares line is different than our eyeballed line;
and we know its the **best line** in a certain sense.

What is a good line?

Can we do better than the eyeball method?

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

That involves

- choosing a criteria, i.e., quantifying how good a line is
- and matching that with a solution, i.e., finding the best line subject to that criteria.

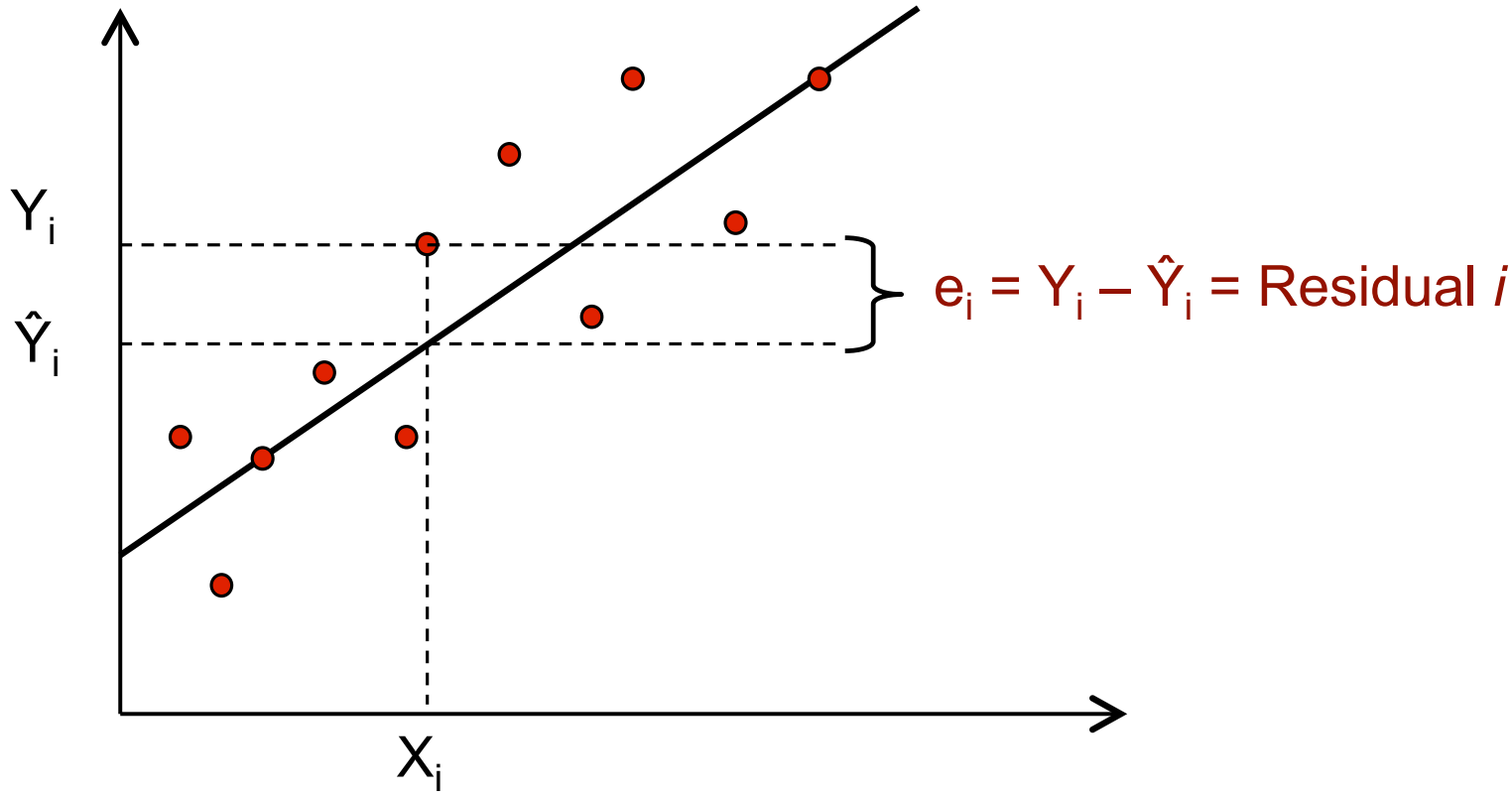
Although there are lots of ways to choose a criteria

- only a small handful lead to solutions that are "easy" to compute,
- and which have nice statistical properties.

Most reasonable criteria involve measuring the amount by which the fitted value obtained from the line differs from the observed value of the response value(s) in the data.

This amount is called the residual.

- Good lines produce small residuals.
- Good lines produce accurate predictions.



The line is our predictions or fitted values : $\hat{Y}_i = b_0 + b_1 X_i$.

The residual e_i is the discrepancy between the fitted $\hat{Y}_i$ and observed Y_i values.

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

Least squares

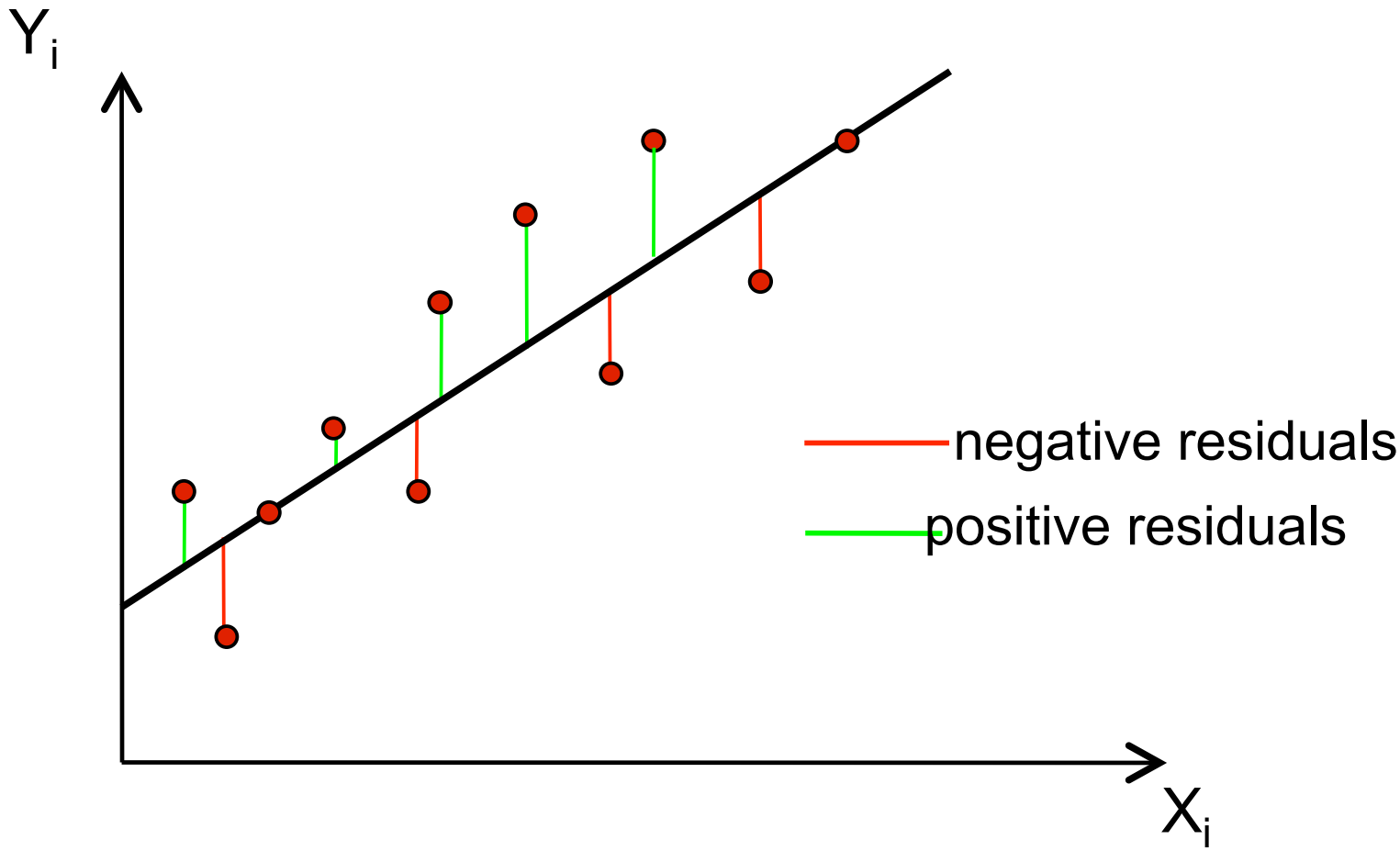
A reasonable goal is to minimize the size of **all** residuals:

- 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.

Since some residuals are positive and some are negative, we need one more ingredient.

- $|e_i|$ treats positives and negatives equally.
- So does e_i^2, which is easier to work with mathematically.

Least squares chooses b_0 and b_1 to minimize $\sum_{i=1}^{n} e_i^2$.



Choose the line to minimize the sum of the squares of the residuals,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - [b_0 + b_1 X_i])^2.$$

Properties of the least squares fit

Developing techniques for model validation and criticism requires a deeper understanding of the least squares line.

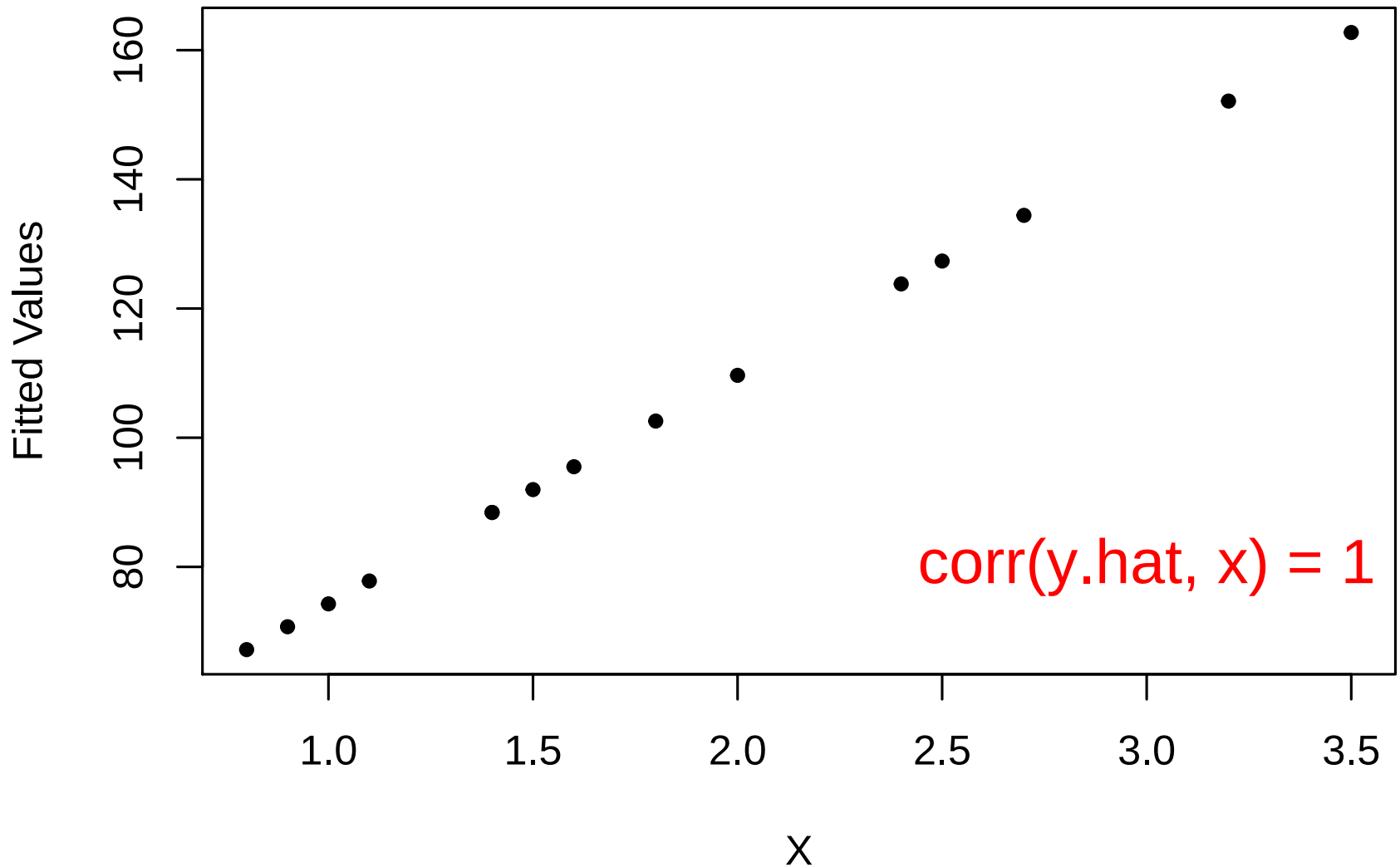
The fitted values ($\hat{Y}_i$) and "residuals" (e_i) obtained from the least squares line have some special properties.

- From now on "obtained from the least squares line" will be implied (and therefore not repeated) whenever we talk about $\hat{Y}_i$ and e_i.

Let's look at the housing data analysis to figure out what some of these properties are ...

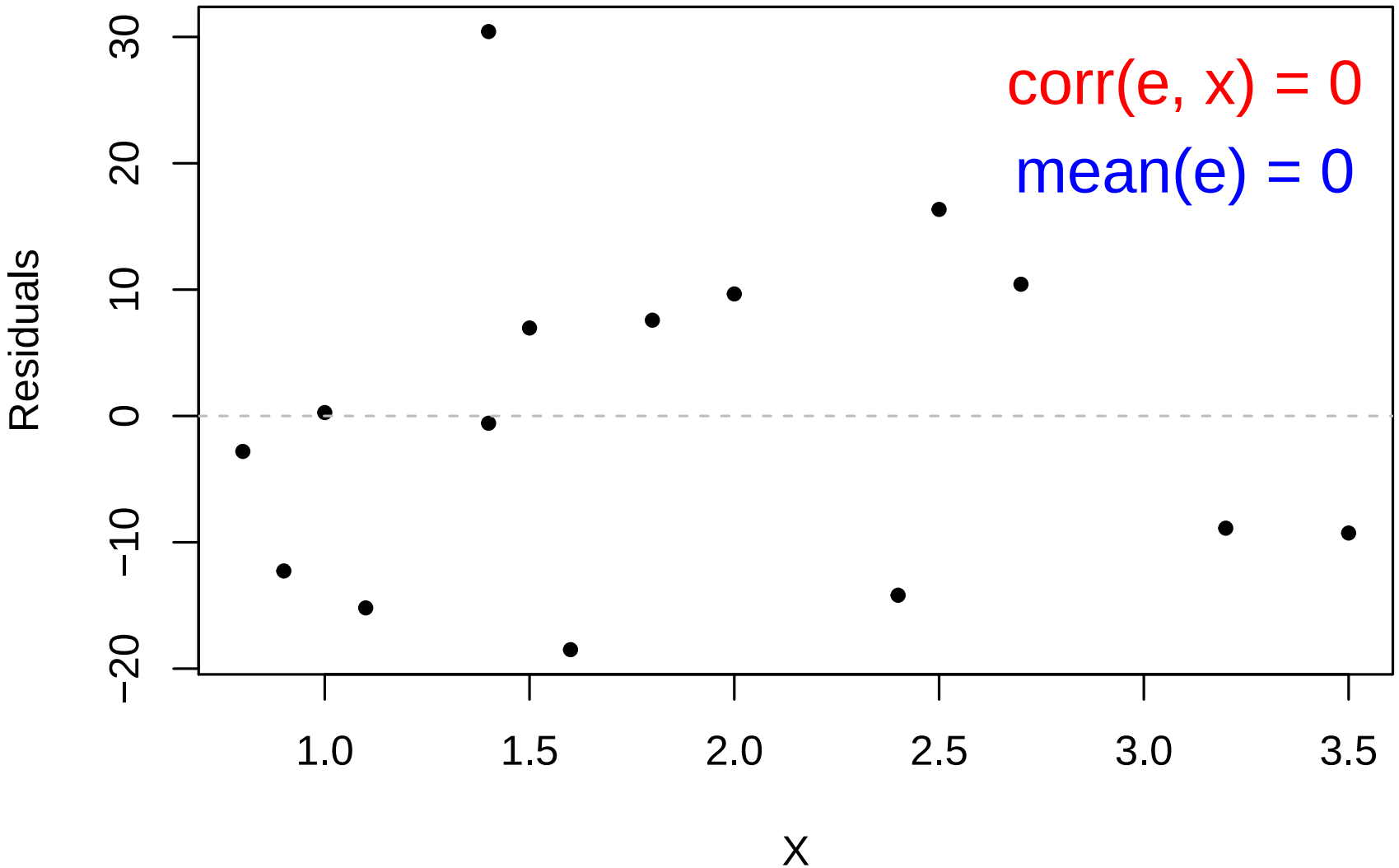
The fitted values are perfectly correlated with the inputs.

```
plot(size, reg$fitted, pch=20, xlab="X", ylab="Fitted Values")
text(x=3, y=80, col=2, cex=1.5,
    paste("corr(y.hat, x) =", cor(size, reg$fitted)))
```



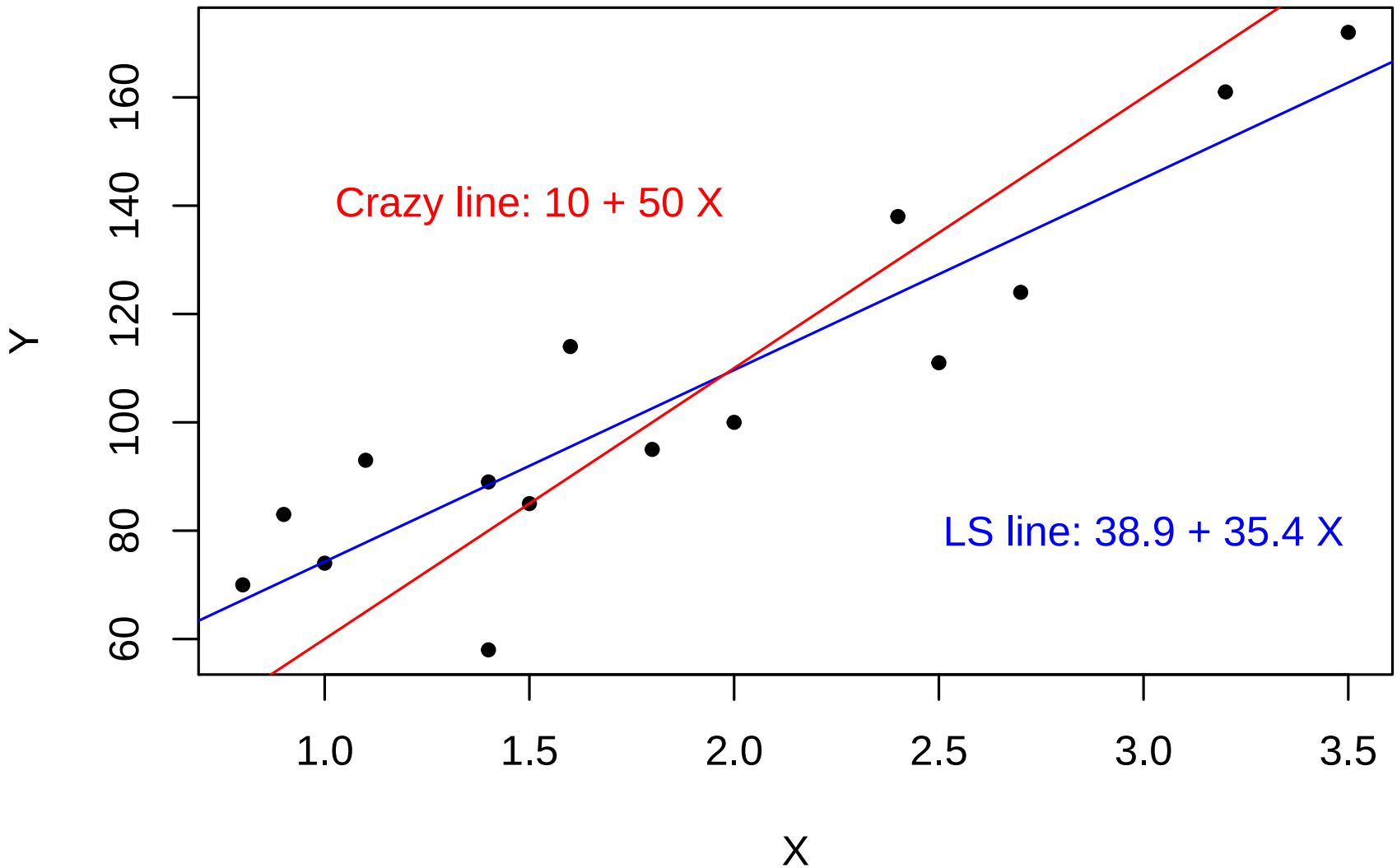
The residuals are “stripped of all linearity”.

```
plot(size, reg$fitted-price, pch=20, xlab="X", ylab="Residuals")
text(x=3.1, y=26, col=2, cex=1.5,
     paste("corr(e, x) =", round(cor(size, reg$fitted-price),2)))
text(x=3.1, y=19, col=4, cex=1.5, paste("mean(e) =", round(mean(reg$fitted-price),0)))
abline(h=0, col=8, lty=2)
```

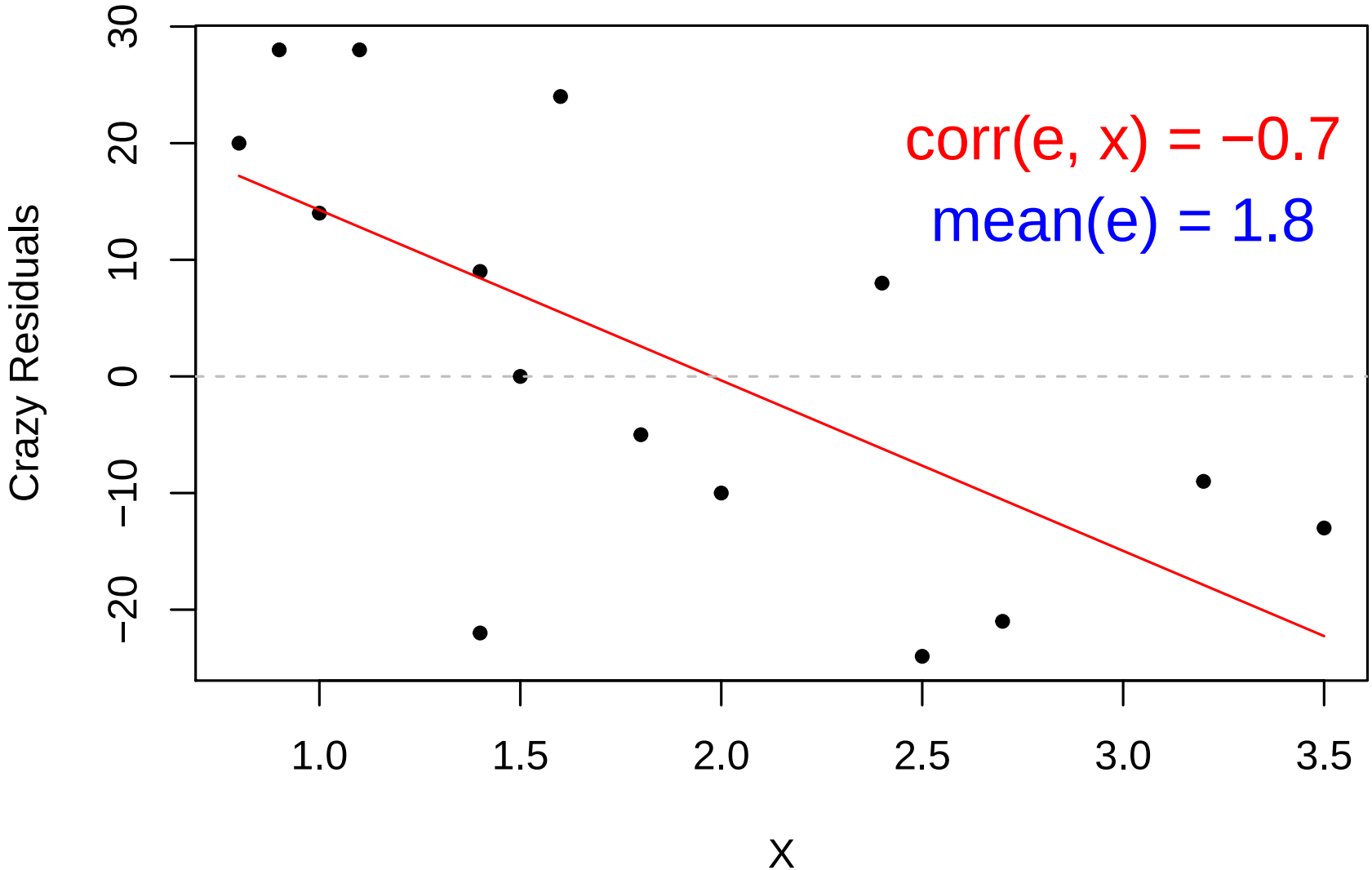


What is the intuition for the relationship between $\hat{Y}$, e and X?

- Lets consider some "crazy" alternative line:



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!

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

$$\min \sum_{i=1}^{n} e_i^2 \quad \text{equivalent to} \quad \text{corr}(e, X) = 0 \ \& \ \frac{1}{n} \sum_{i=1}^{n} e_i = 0$$

We need to exploit all of the (linear!) 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"; $\text{corr}(X, \hat{Y}) = 1$;
- e is unrelated to X; $\text{corr}(X, e) = 0$.

To summarize:

R's `lm(Y ~ X)` function

- finds the coefficients b_0 and b_1 characterizing the "least squares" line $\hat{Y} = b_0 + b_1 X$.
- That is it minimizes $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$.
- Equivalent to: $\text{corr}(e, X) = 0$ & $\frac{1}{n}\sum_{i=1}^{n} e_i = 0$

The least squares formulas are

$$b_1 = r_{xy}\frac{s_y}{s_x} \quad \text{and} \quad b_0 = \bar{Y} - b_1\bar{X}.$$

Glossary of symbols

- X = input, explanatory variable, or covariate.
- Y = output, dependent variable, or response.
- s_{xy} is covariance and r_{xy} is the correlation, s_x and s_y are standard deviation of X and Y respectively
- $r_{xy} = \frac{s_{xy}}{s_x s_y}$.
- b_0 = least squares estimate of the intercept
- b_1 = least squares estimate of the slope
- $\hat{Y}$ is the fitted value $b_0 + b_1 X$
- e_i is the residual $Y_i - \hat{Y}_i$.