



McCOMBS SCHOOL OF BUSINESS

**Salem Center for Policy**

# Prediction

David Puelz

October 5, 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.



**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

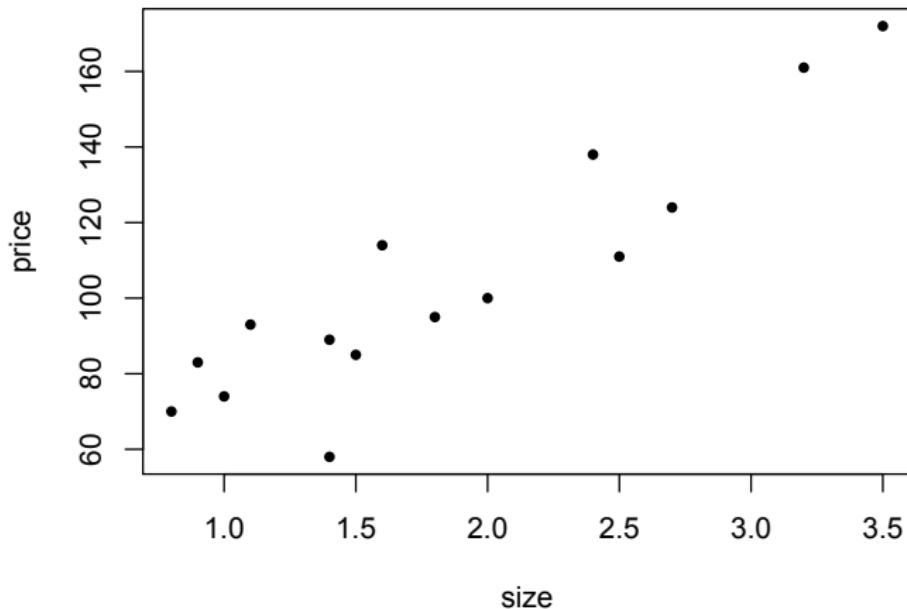


What does this data look like?

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

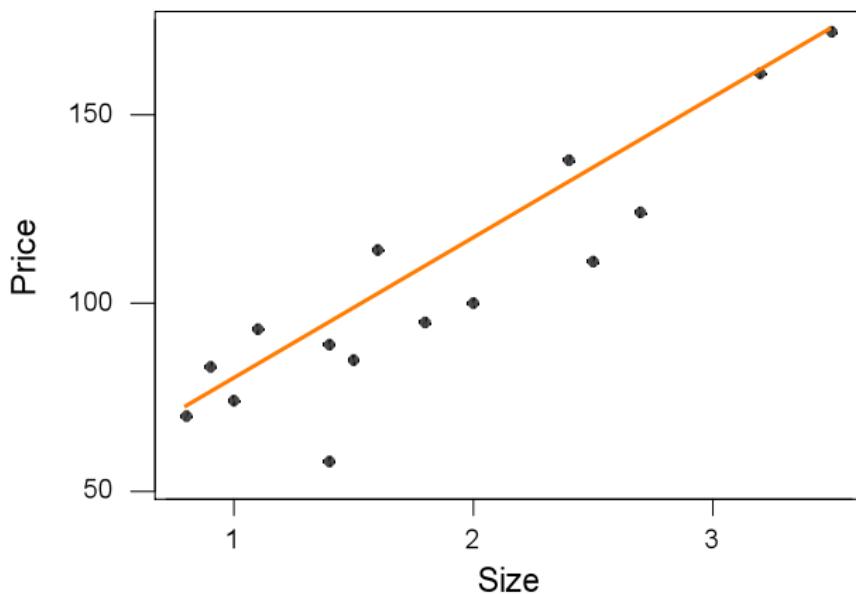
A linear relationship is written

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

# Linear prediction

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

As size goes up, price goes up.



## Linear prediction



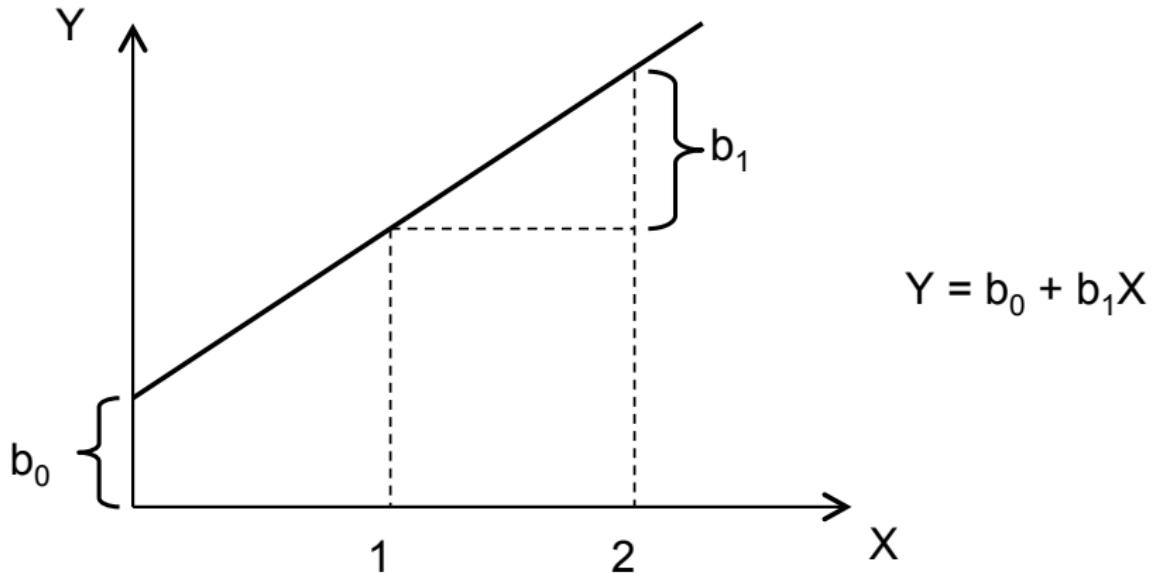
Recall that the equation of a line is:

$$Y = b_0 + b_1 X$$

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





## Q: How to find the “best line”?

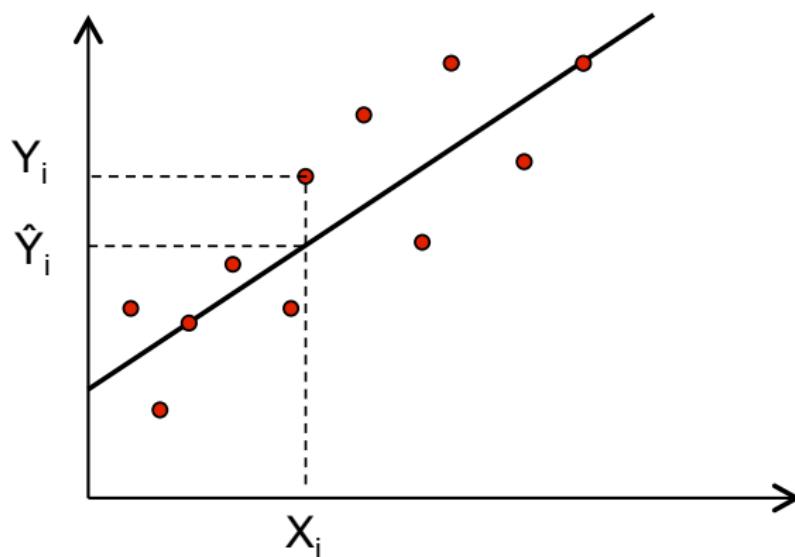
We desire a strategy for estimating the slope and intercept parameters in the model  $\hat{Y} = b_0 + b_1 X$

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

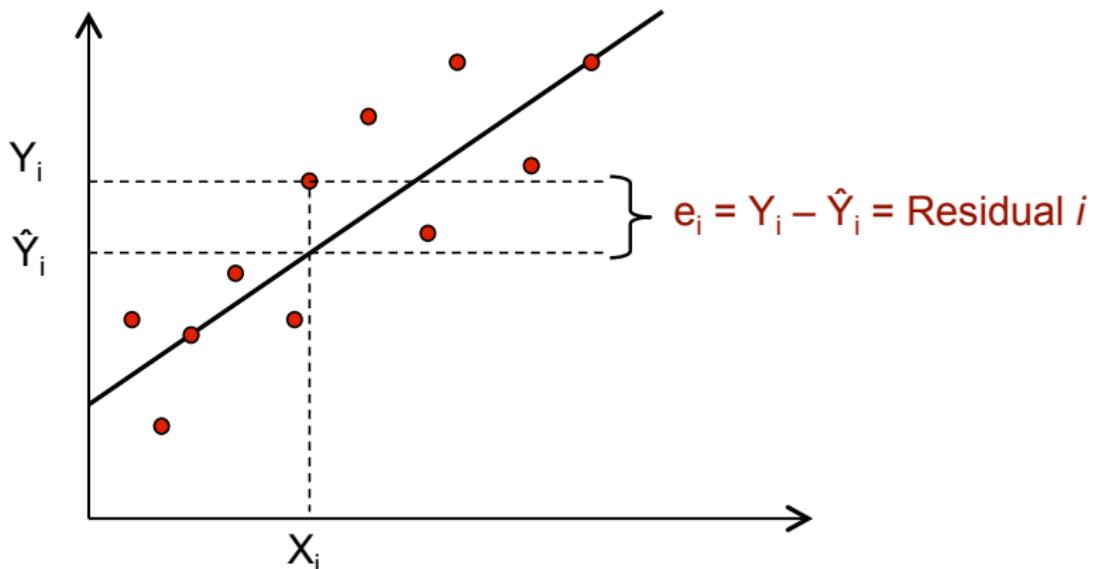
What is the “fitted value”?



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

# Linear prediction

What is the “residual” for the  $i$ th observation?



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



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.



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  $\sum_{i=1}^N e_i^2$

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



## 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



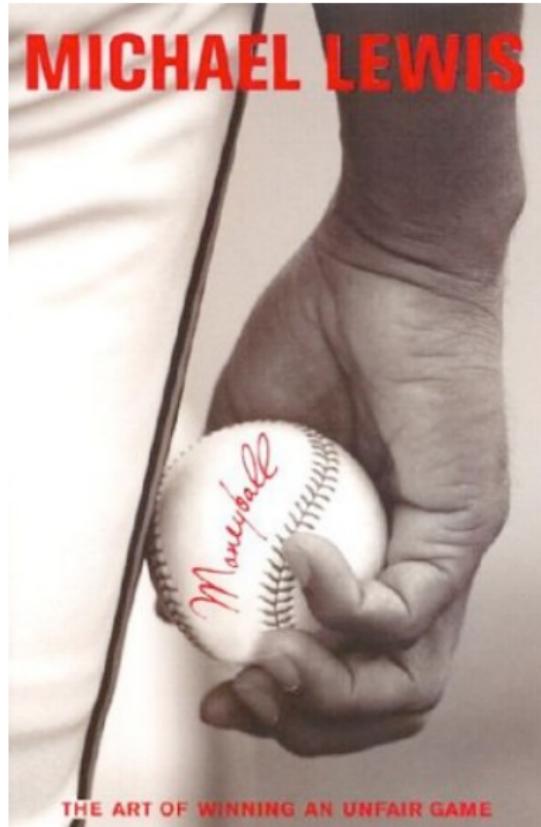
### Problems:

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

### Solutions:

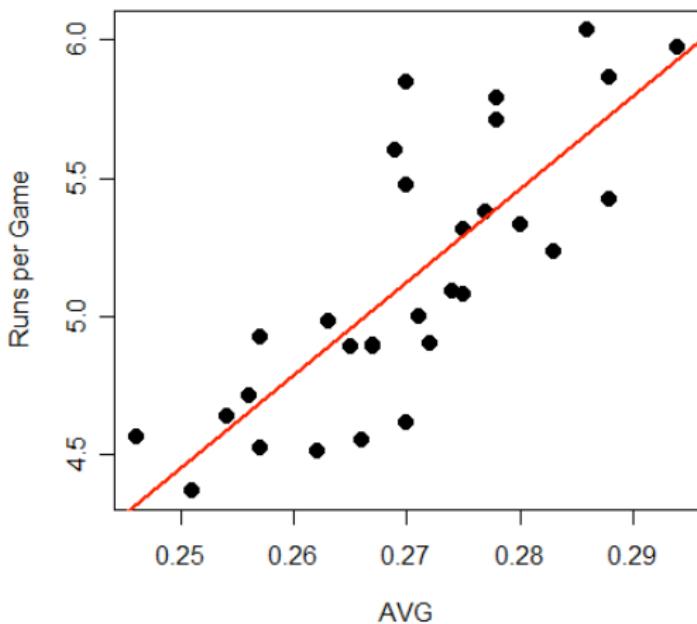
- Compare *prediction rules* that forecast runs as a function of either **AVG** (batting average), **SLG** (slugging percentage – total bases divided by at bats) or **OBP** (on base percentage)

## Example 2: Offensive performance in baseball



# Baseball data – using AVG

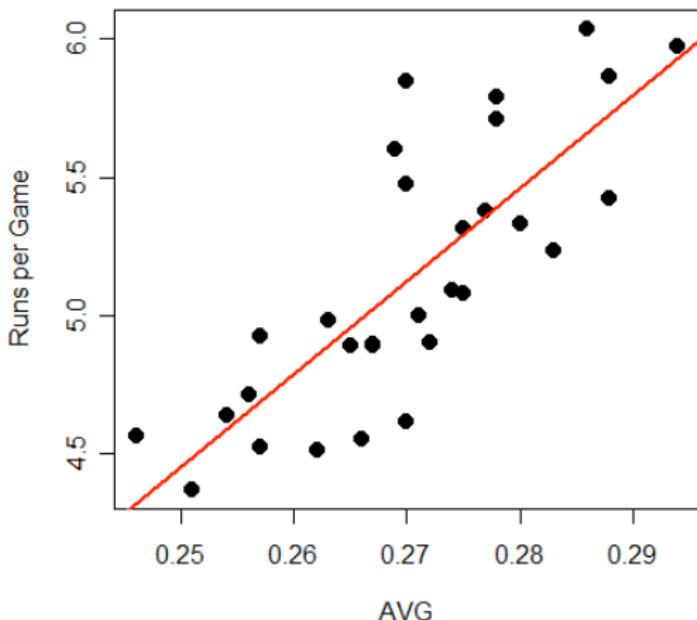
Each observation corresponds to a team in MLB. Each quantity is the average over a season.



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

LS fit:  $\text{Runs/Game} = -3.93 + 33.57 \text{ AVG}$

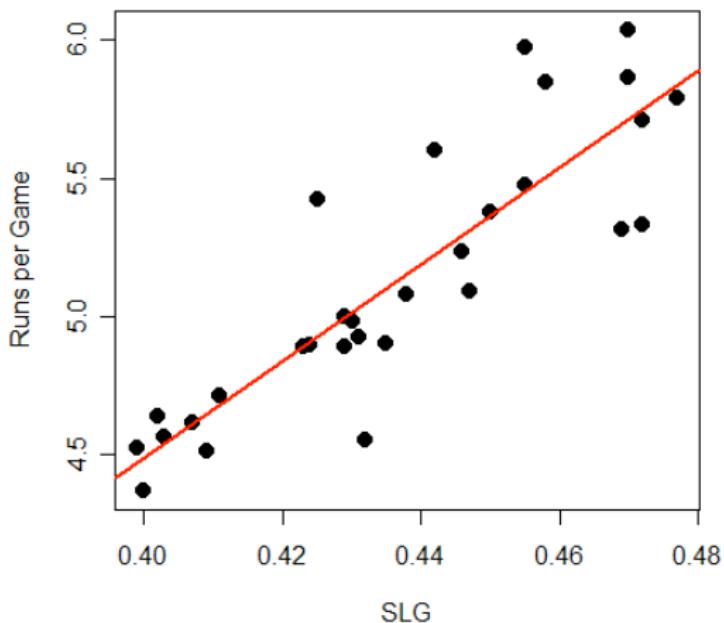
# Baseball data – using AVG



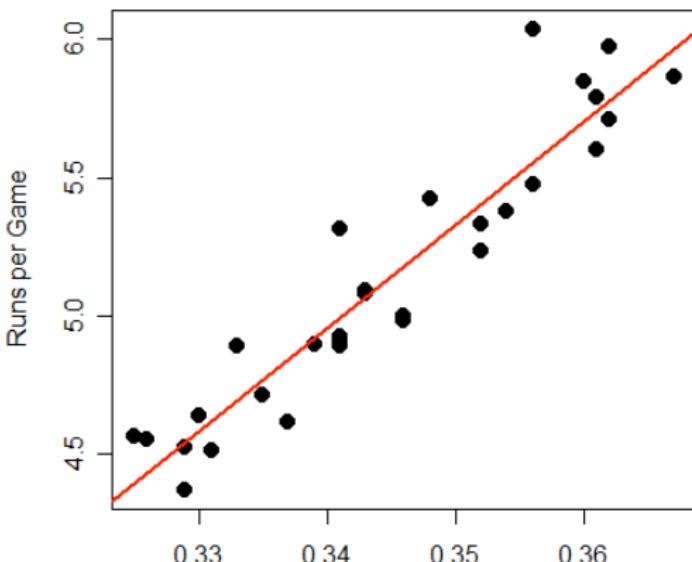
$Y$  = runs per game;  $X$  = AVG (average)

LS fit:  $\text{Runs/Game} = -3.93 + 33.57 \text{ AVG}$

# Baseball Data – using SLG



# Baseball Data – using OBP



$Y = \text{runs per game}; X = \text{OBP}$  (on base percentage)  
LS fit:  $\text{Runs/Game} = -7.78 + 37.46 \text{ OBP}$



- What is the best prediction rule?
- Let's compare the predictive ability of each model using the average squared error

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

Place your money on OBP!!!



Root Mean Squared Error	
AVG	0.29
SLG	0.23
OBP	0.16



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:

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

where,

- $\bar{X}$  and  $\bar{Y}$  are the sample mean of  $X$  and  $Y$
- $\text{corr}(x, y) = r_{xy}$  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**

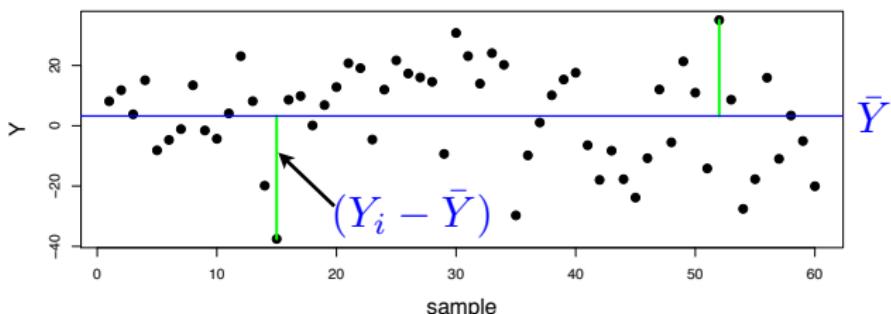
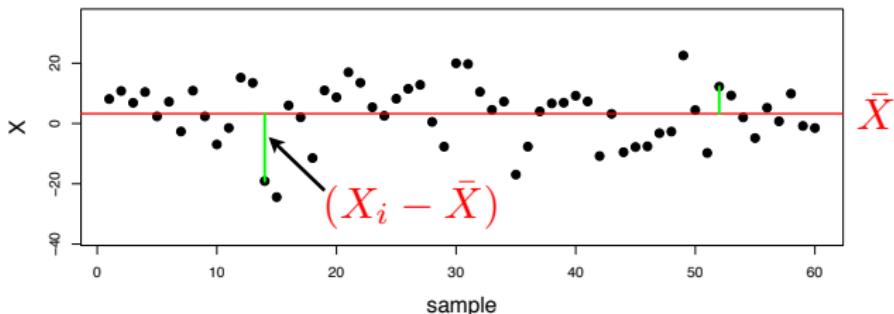
$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

- Sample Standard Deviation:

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

# Visual: standard deviation

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

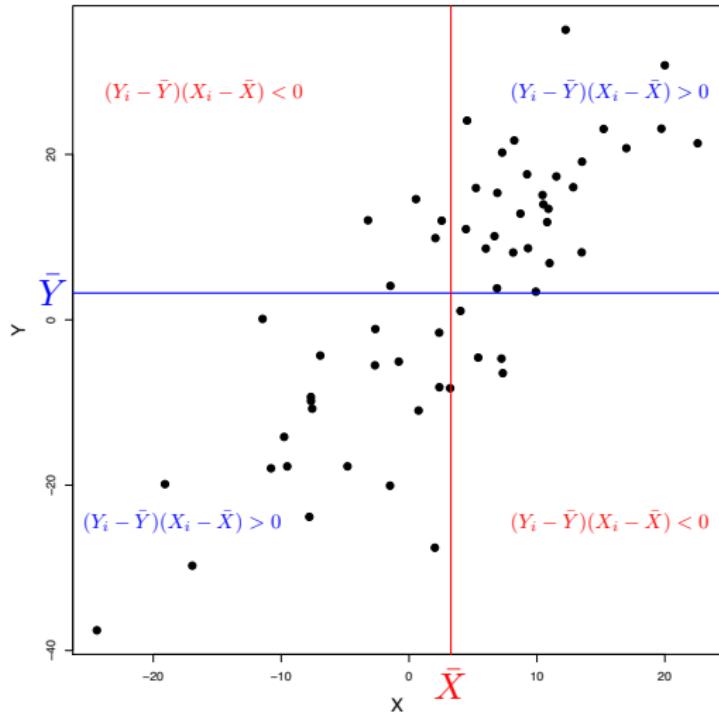


$$s_x = 9.7 \quad s_y = 15.98$$

# 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:

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

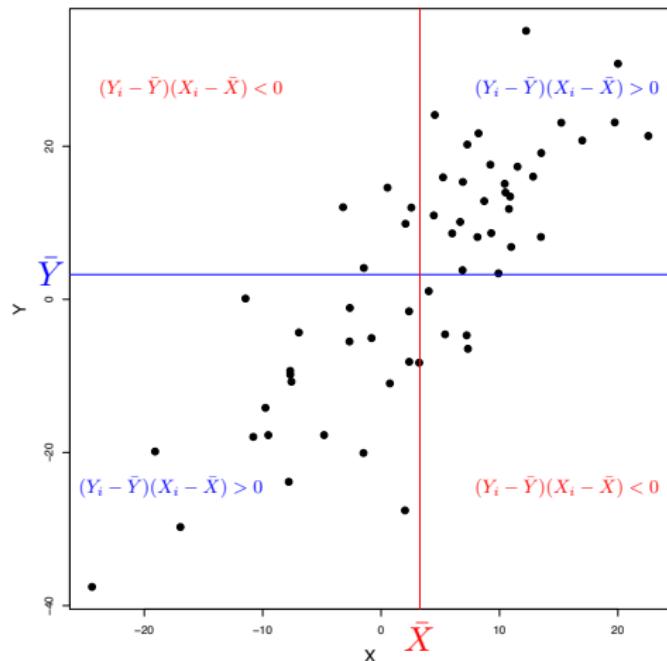
The correlation is scale invariant and the units of measurement don't matter: It is always true that  $-1 \leq \text{corr}(X, Y) \leq 1$ .

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

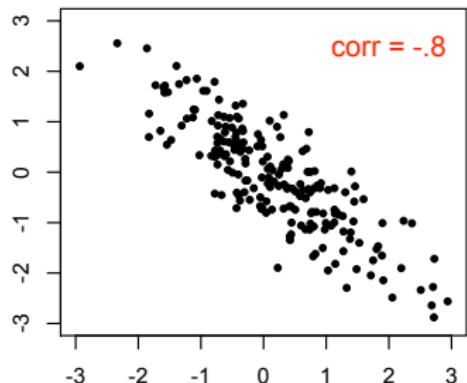
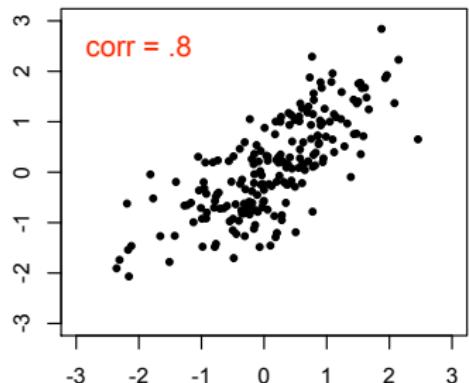
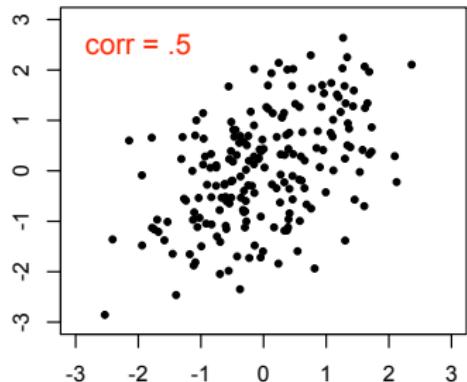
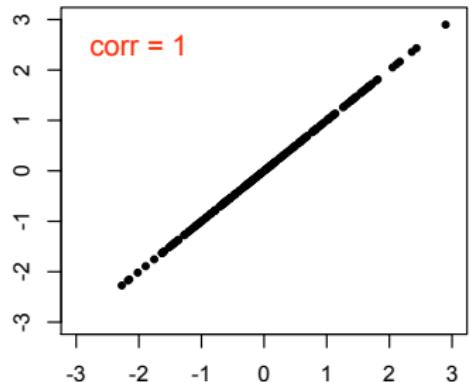
# Correlation



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



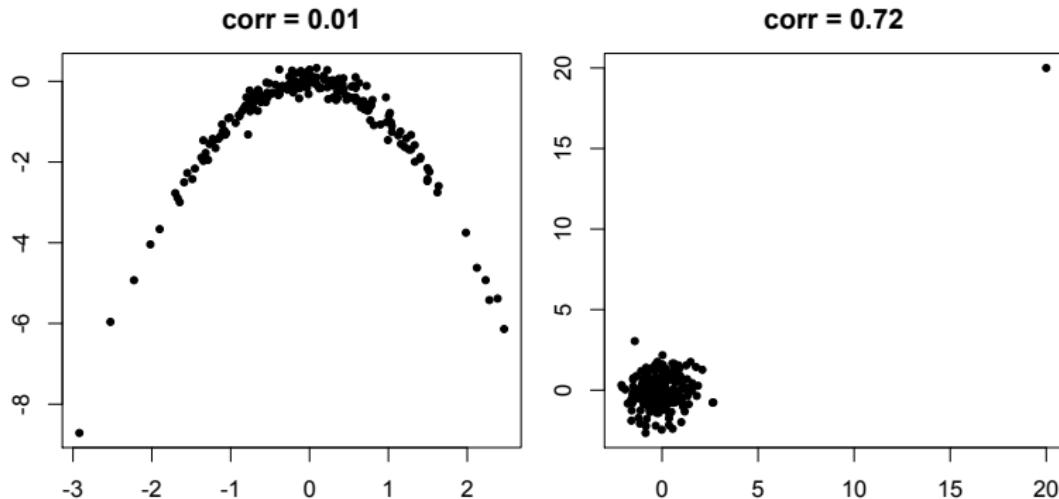
# Correlation



# 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:

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

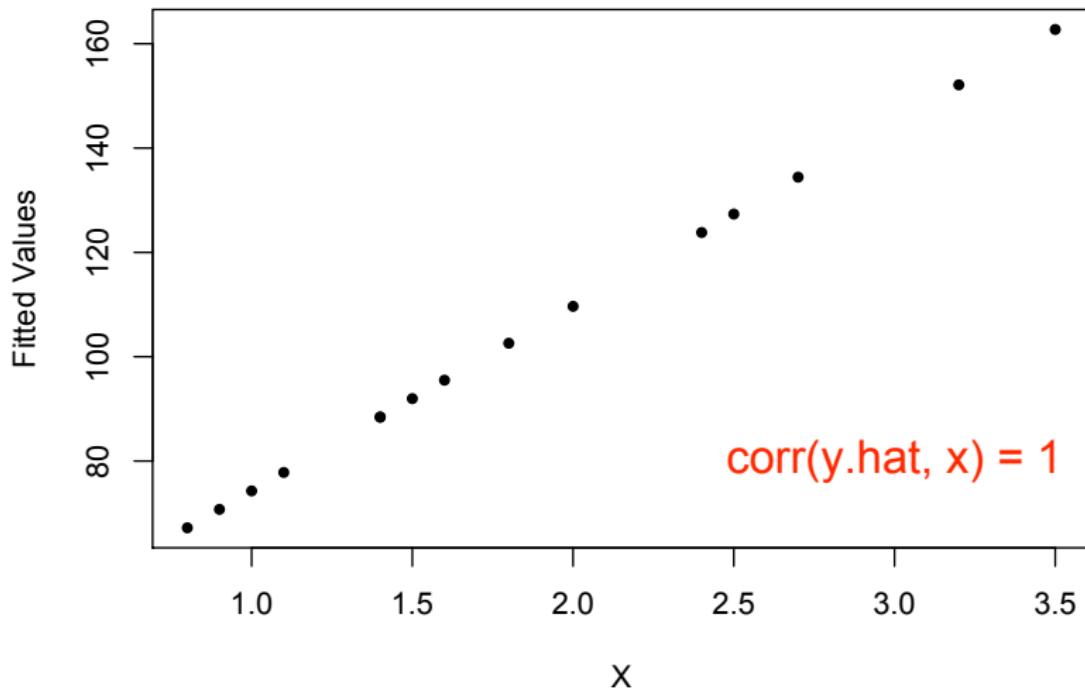
So, the right slope is the **correlation coefficient** times a **scaling factor** that ensures the proper units for  $b_1$



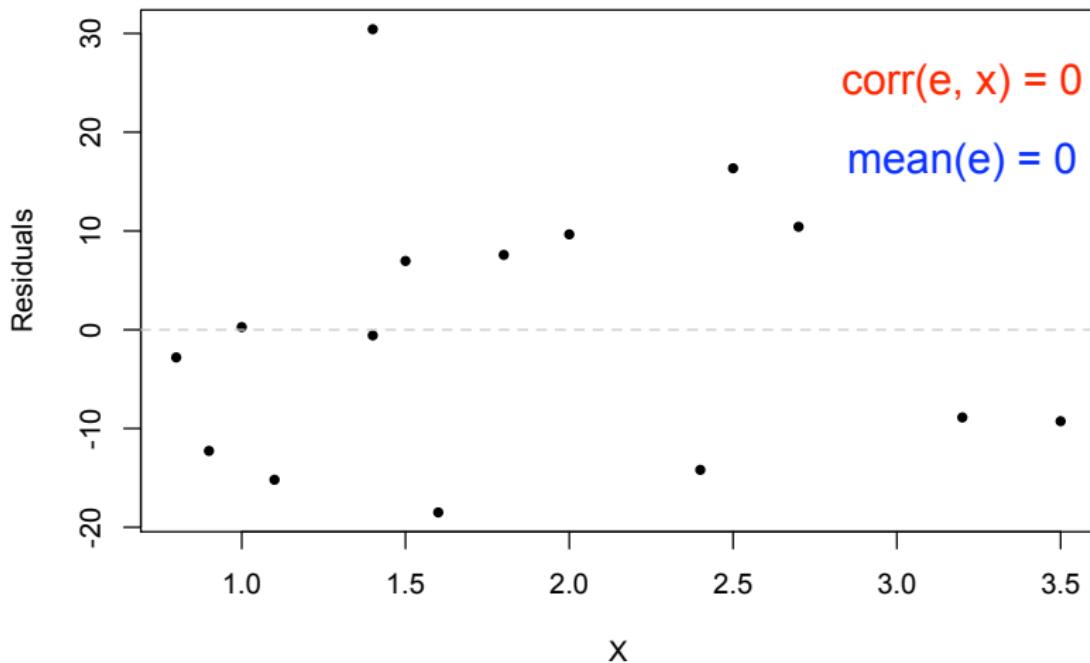
From now on, terms “fitted values” ( $\hat{Y}_i$ ) and “residuals” ( $e_i$ ) refer to those obtained from the least squares line.

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

# The fitted values and X



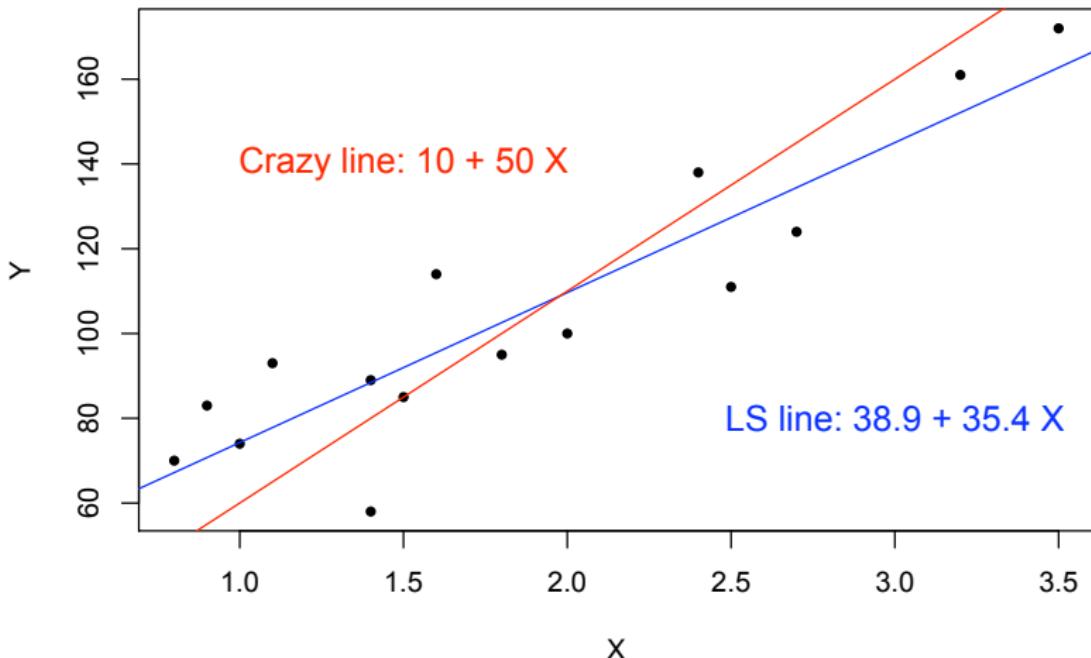
# The residuals and X



# Why?



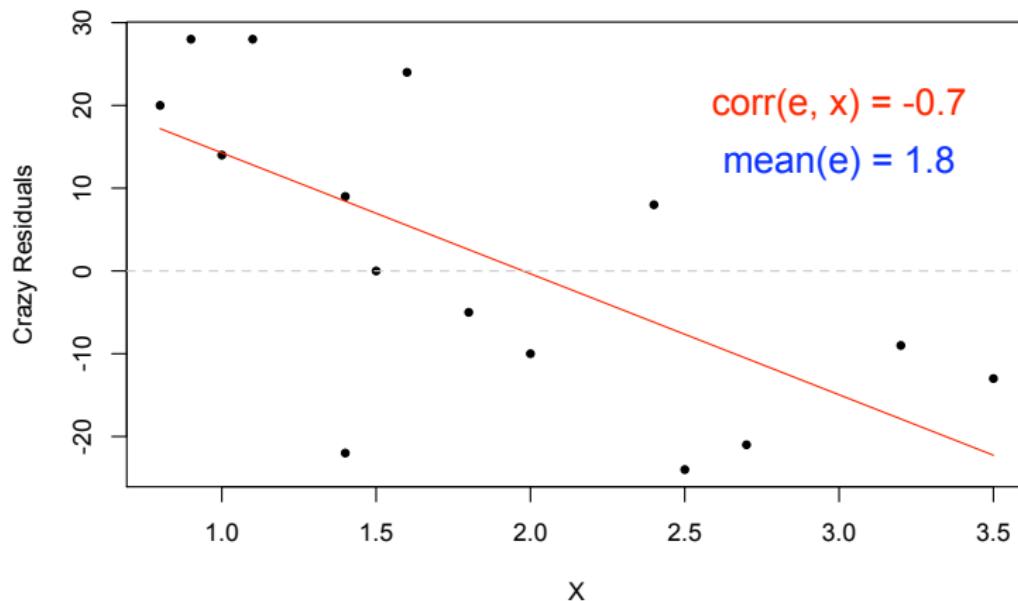
What is the intuition for the relationship between  $\hat{Y}$  and  $e$  and  $X$ ?  
Let's consider some "crazy" alternative line:



## 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 “ $X$ ness” 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  $X$ ;  $\text{corr}(X, e) = 0$ .

# Decomposing the variance

Q: How well does the least squares line explain variation in  $Y$ ?

Remember that  $Y = \hat{Y} + e$

Since  $\hat{Y}$  and  $e$  are uncorrelated, i.e.  $\text{corr}(\hat{Y}, e) = 0$ ,

$$\text{var}(Y) = \text{var}(\hat{Y} + e) = \text{var}(\hat{Y}) + \text{var}(e)$$

$$\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2}{n-1} + \frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n-1}$$

Given that  $\bar{e} = 0$ , and  $\bar{\hat{Y}} = \bar{Y}$  (why?) we get to:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n e_i^2$$

# Decomposing the variance

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n e_i^2$$

The equation is decomposed into three parts using curly braces under each term:

- $\sum_{i=1}^n (Y_i - \bar{Y})^2$  is bracketed and labeled "Total Sum of Squares SST".
- $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$  is bracketed and labeled "Regression SS SSR".
- $\sum_{i=1}^n e_i^2$  is bracketed and labeled "Error SS SSE".

SSR: Variation in  $Y$  explained by the regression line.

SSE: Variation in  $Y$  that is left unexplained.

## Decomposing the variance

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n e_i^2$$

Total Sum of Squares  
SST

Regression SS  
SSR

Error SS  
SSE

SSR: Variation in  $Y$  explained by the regression line.

SSE: Variation in  $Y$  that is left unexplained.

$$\text{SSR} = \text{SST} \Rightarrow \text{perfect fit.}$$

*Be careful of similar acronyms; e.g. SSR for “residual” SS.*

## A goodness of fit measure: $R^2$



The coefficient of determination, denoted by  $R^2$ , measures goodness of fit:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

- $0 < R^2 < 1$ .
- The closer  $R^2$  is to 1, the better the fit.

## Back to baseball



Three very similar, related ways to look at a simple linear regression... with only one  $X$  variable, life is easy!

	$R^2$	corr	SSE
OBP	0.88	0.94	0.79
SLG	0.76	0.87	1.64
AVG	0.63	0.79	2.49



Prediction and regression + probability

## Prediction and the modeling goal



A prediction rule is any function where you input  $X$  and it outputs  $\hat{Y}$  as a predicted response at  $X$ .

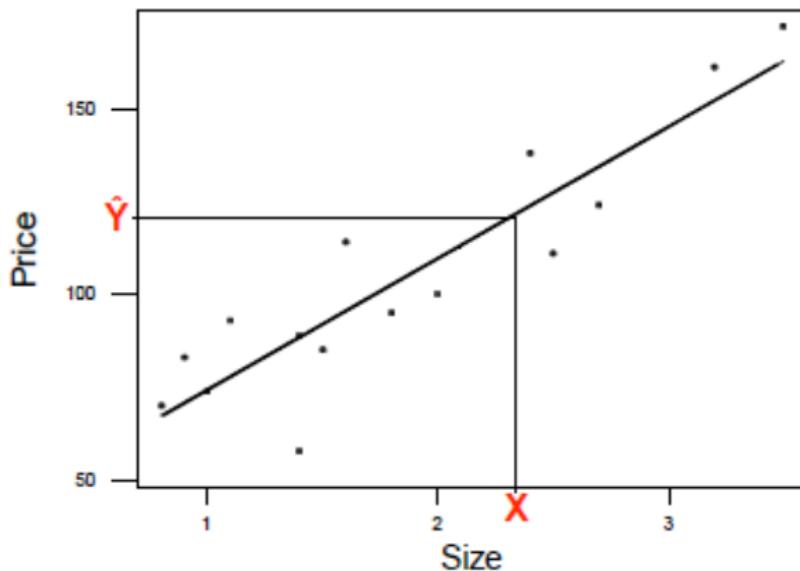
The least squares line is a prediction rule:

$$\hat{Y} = f(X) = b_0 + b_1 X$$

# Prediction and the modeling goal

$\hat{Y}$  is not going to be a perfect prediction.

We need to devise a notion of **forecast accuracy**.





# Prediction and the modeling goal

There are two things that we want to know:

- What value of  $Y$  can we expect for a given  $X$ ?
- How sure are we about this forecast? Or how different could  $Y$  be from what we expect?

Our goal is to measure the accuracy of our forecasts or **how much uncertainty there is in the forecast**. One method is to specify a range of  $Y$  values that are likely, given an  $X$  value.

**Prediction Interval: probable range for  $Y$ -values given  $X$**

# Prediction and the modeling goal



**Key Insight:** To construct a prediction interval, we will have to assess the likely range of error values corresponding to a  $Y$  value that has not yet been observed!

We will build a **probability model** (e.g., Normal distribution).

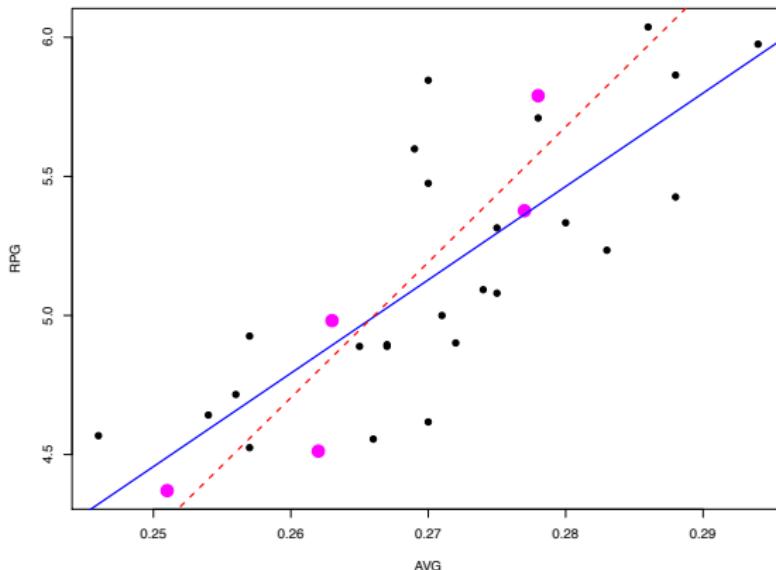
Then we can say something like “with 95% probability the error will be no less than -\$28,000 or larger than \$28,000”.

We must also acknowledge that the “fitted” line may be fooled by particular realizations of the residuals.

# Prediction and the modeling goal



We are always looking at samples! The dashed line fits the purple points. The solid line fits all the points. Which line is better? Why?



In summary, we need to work with the notion of a “true line” and a probability distribution that describes deviation around the line.

# The simple linear regression model

The power of statistical inference comes from the ability to make precise statements about the accuracy of the forecasts.

In order to do this we must invest in a **probability model**.

Simple Linear Regression Model:  $Y = \beta_0 + \beta_1 X + \varepsilon$

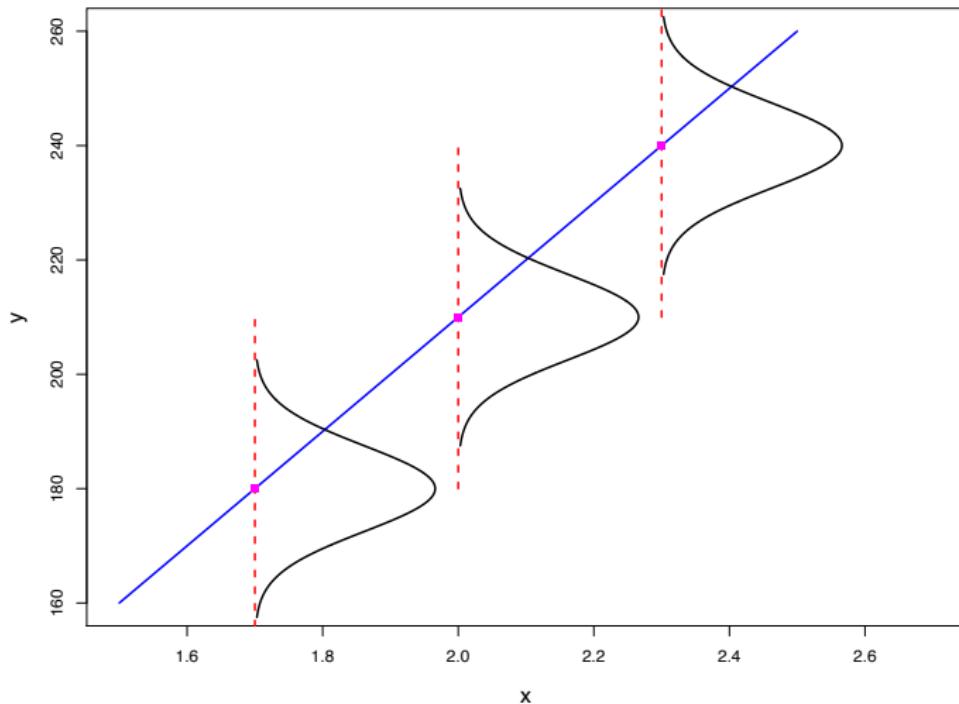
$$\varepsilon \sim N(0, \sigma^2)$$

- $\beta_0 + \beta_1 X$  represents the “true line”; The part of  $Y$  that depends on  $X$ .
- The error term  $\varepsilon$  is independent “idiosyncratic noise”; The part of  $Y$  not associated with  $X$ .

# The Simple Linear Regression Model



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



## The simple linear regression model – example



You are told (without looking at the data) that

$$\beta_0 = 40; \beta_1 = 45; \sigma = 10$$

and you are asked to predict price of a 1500 square foot house.

What do you know about  $Y$  from the model?

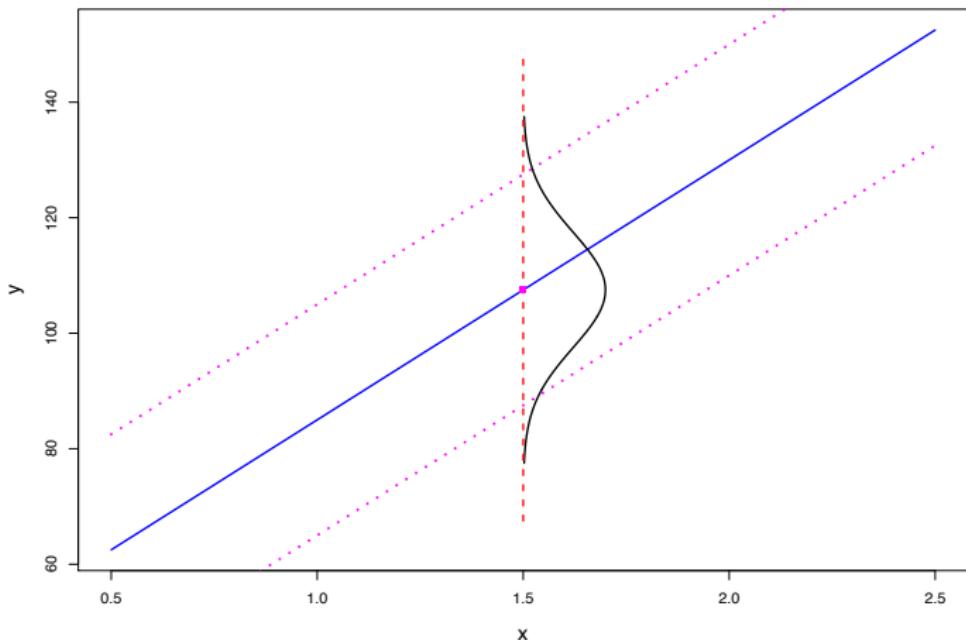
$$\begin{aligned} Y &= 40 + 45(1.5) + \varepsilon \\ &= 107.5 + \varepsilon \end{aligned}$$

Thus our prediction for price is  $Y|X = 1.5 \sim N(107.5, 10^2)$   
and a 95% Prediction Interval for  $Y$  is  $87.5 < Y < 127.5$

# Conditional distributions



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



The conditional distribution for  $Y$  given  $X$  is Normal:

$$Y|X = x \sim N(\beta_0 + \beta_1 x, \sigma^2).$$

# Conditional distributions



The model says that the mean value of a 1500 sq. ft. house is \$107,500 and that deviation from mean is within  $\approx \$20,000$ .

We are 95% sure that

- $-20 < \varepsilon < 20$
- $87.5 < Y < 127.5$

In general, the 95 % Prediction Interval is  $PI = \beta_0 + \beta_1 X \pm 2\sigma$ .

# Why do we choose this probability model?



Put differently, why do we have  $\varepsilon \sim N(0, \sigma^2)$ ?

- $E[\varepsilon] = 0 \Leftrightarrow E[Y | X] = \beta_0 + \beta_1 X$   
( $E[Y | X]$  is “conditional expectation of  $Y$  given  $X$ ”).
- Many things are close to Normal (central limit theorem).
- It works! This is a very robust model for the world.

We can think of  $\beta_0 + \beta_1 X$  as the “true” regression line.

# Conditional distributions

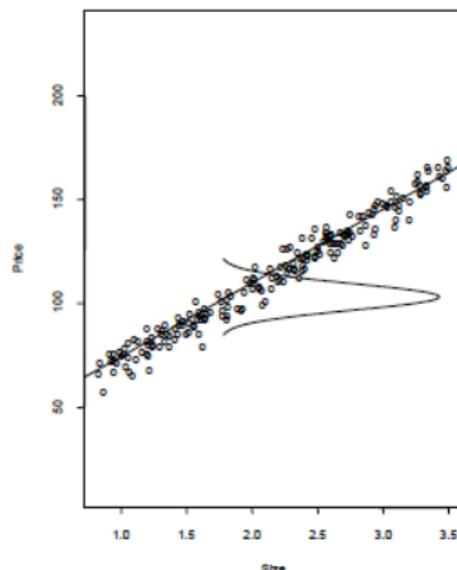


The conditional distribution for  $Y$  given  $X$  is Normal:

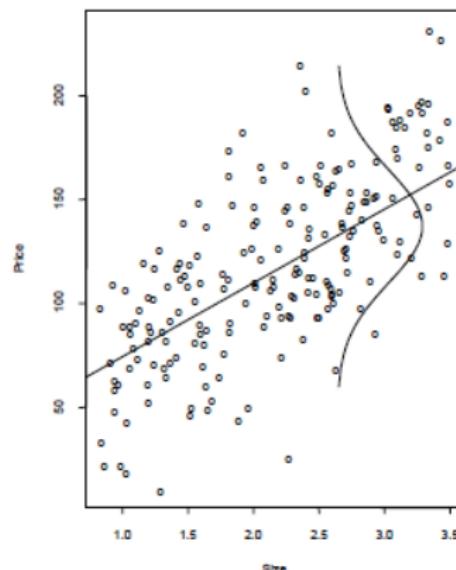
$$Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2).$$

$\sigma$  controls dispersion:

$\sigma$  small /  $\varepsilon$  small



$\sigma$  large /  $\varepsilon$  large



# Conditional distributions



More on the conditional distribution:

$$Y|X \sim N(E[Y|X], \text{var}(Y|X)).$$

- The conditional mean is

$$E[Y|X] = E[\beta_0 + \beta_1 X + \varepsilon] = \beta_0 + \beta_1 X.$$

- The conditional variance is

$$\text{var}(Y|X) = \text{var}(\beta_0 + \beta_1 X + \varepsilon) = \text{var}(\varepsilon) = \sigma^2.$$

- $\sigma^2 < \text{var}(Y)$  if  $X$  and  $Y$  are related.

## Summary of simple linear regression



Assume that all observations are drawn from our regression model and that errors on those observations are independent.

The model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where  $\varepsilon$  is independent and identically distributed  $N(0, \sigma^2)$ .

- **independence** means that knowing  $\varepsilon_i$  doesn't affect your views about  $\varepsilon_j$
- **identically distributed** means that we are using the same Normal for every  $\varepsilon_i$

# Summary of simple linear regression



The model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2).$$

The SLR has 3 basic parameters:

- $\beta_0, \beta_1$  (linear pattern)
- $\sigma$  (variation around the line).

# Key characteristics of linear regression model



- Mean of  $Y$  is **linear** in  $X$ .
- Error terms (deviations from line) are **Normally distributed** (very few deviations are more than 2 standard deviations away from the regression mean).
- Error terms have **constant variance**.

# Estimation for the SLR model

SLR assumes every observation in the dataset was generated by the model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

This is a model for the conditional distribution of Y given X.

We use Least Squares *to estimate*  $\beta_0$  and  $\beta_1$ :

$$\hat{\beta}_1 = b_1 = r_{xy} \times \frac{s_y}{s_x}$$

$$\hat{\beta}_0 = b_0 = \bar{Y} - b_1 \bar{X}$$

## Estimation of error variance



We estimate  $s^2$  with:

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2 = \frac{SSE}{n-2}$$

(2 is the number of regression coefficients; i.e. 2 for  $\beta_0$  and  $\beta_1$ ).

We have  $n - 2$  degrees of freedom because 2 have been “used up” in the estimation of  $b_0$  and  $b_1$ .

We usually use  $s = \sqrt{SSE/(n-2)}$ , in the same units as  $Y$ . It's also called the **regression standard error**.



We now know how to make statements about **uncertainty** in forecasts aka predictions. But what about our **parameter estimates**?

**Q:** How much do our estimates depend on the particular random sample that we happen to observe?

# Uncertainty in estimates



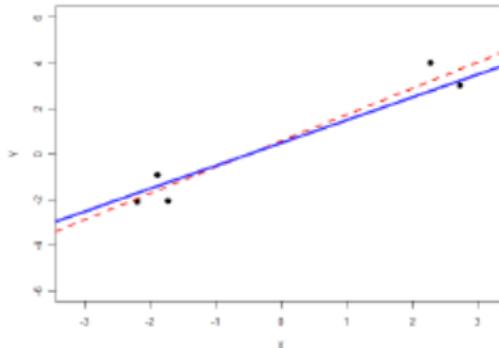
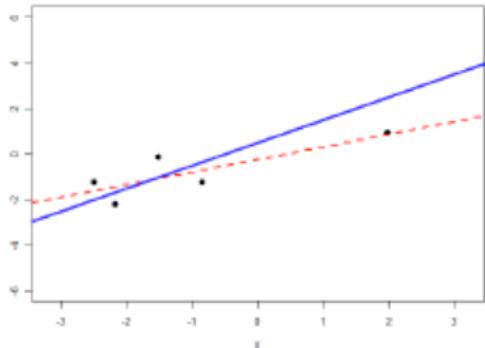
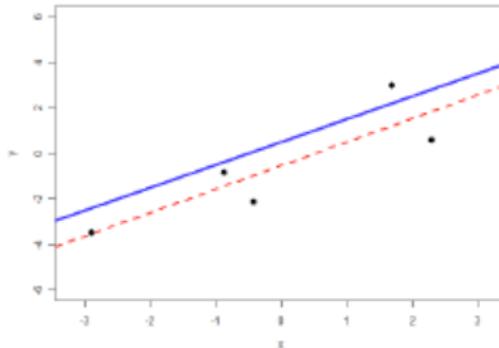
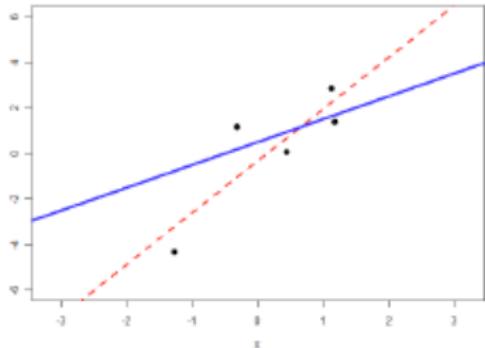
Imagine:

- Randomly draw different samples of the same size.
- For each sample, compute the estimates  $b_0$ ,  $b_1$ , and  $s$ .

If the estimates don't vary much from sample to sample, then it doesn't matter which sample you happen to observe.

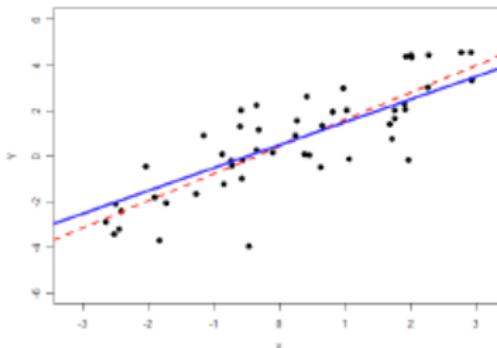
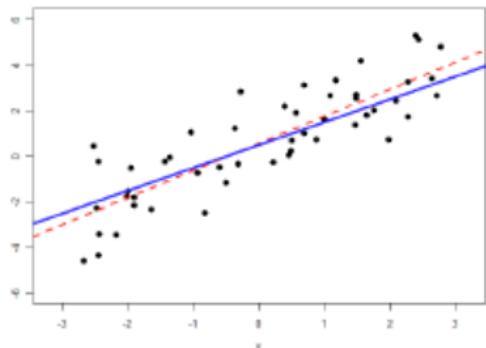
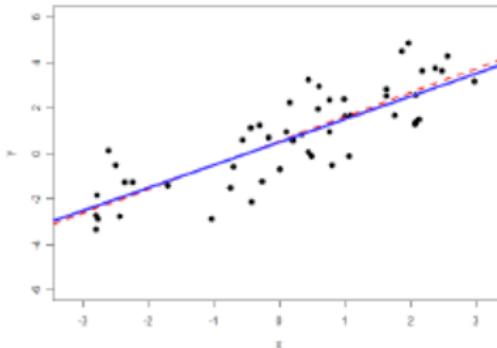
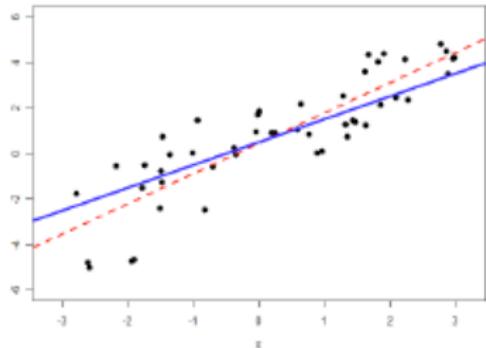
If the estimates do vary a lot, then it matters which sample you happen to observe.

# Sampling distribution of least squares estimates



N=5  
True Model ———  
LS Line -----

# Sampling distribution of least squares estimates



N=50  
True Model —  
LS Line - - -



Least squares lines are much closer to the true line when  $N = 50$ .

For  $N = 5$ , some lines are close, others aren't:

We need to get “lucky!”

So, to sum up ...

Parameter estimates are **uncertain** because we only have sample from the population

We can characterize a **sampling distribution** of each parameter based on the assumptions of the SLR model

The standard deviation of the sampling distribution is called the **standard error**

→ R directly reports standard errors. They can also be calculated with other methods, like bootstrapping.



## 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
```



## The bootstrap

# Quantifying uncertainty



In data science, we equate **trustworthiness** with **stability**:

**Key question:** If our data had been different merely due to chance, would our answer have been different, too? Or would the answer have been stable, even with different data?

Confidence in your estimates  $\iff$  Stability of those estimates under the influence of chance

## Example: Quantifying uncertainty



For example:

- If doctors had taken a different sample of 503 cancer patients and gotten a drastically different estimate of the new treatment's effect, then the original estimate isn't very trustworthy.
- If, on the other hand, pretty much any sample of 503 patients would have led to the same estimates, then their answer for this particular subset of 503 is probably accurate.



## Some notation

Suppose we are trying to estimate some population-level feature of interest,  $\theta$ . This might be something very complicated!

So we take a sample from the population:  $X_1, X_2, \dots, X_N$ . We use the data to form an estimate  $\hat{\theta}_N$  of the parameter.

**Key insight:**  $\hat{\theta}_N$  is a random variable.

( $\hat{\theta}_N$  can be the slope of a least squares regression)



## Some notation

Suppose we are trying to estimate some population-level feature of interest,  $\theta$ . This might be something very complicated!

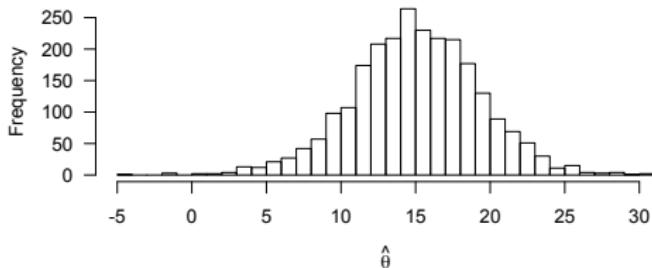
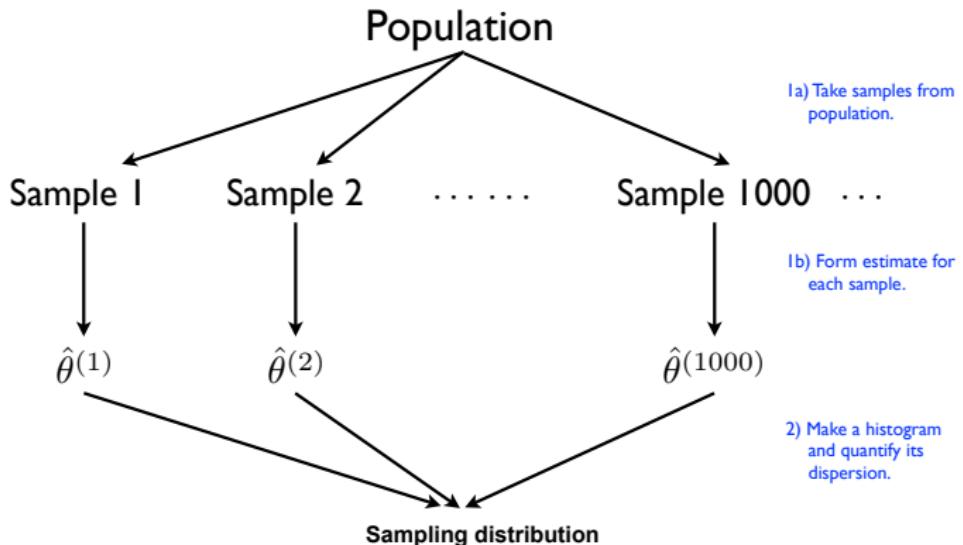
So we take a sample from the population:  $X_1, X_2, \dots, X_N$ . We use the data to form an estimate  $\hat{\theta}_N$  of the parameter.

**Key insight:**  $\hat{\theta}_N$  is a random variable.

( $\hat{\theta}_N$  can be the slope of a least squares regression)

→ Now imagine repeating this process thousands of times! Since  $\hat{\theta}_N$  is a random variable, it has a probability distribution: the sampling distribution.

# Visualizing this procedure



## Revisiting the standard error



**Standard error:** the standard deviation of an estimator's sampling distribution:

$$\begin{aligned}\text{se}(\hat{\theta}_N) &= \sqrt{\text{var}(\hat{\theta}_N)} \\ &= \sqrt{E[(\hat{\theta}_N - \bar{\theta}_N)^2]} \\ &= \text{Typical deviation of } \hat{\theta}_N \text{ from its average}\end{aligned}$$

"If I were to take repeated samples from the population and use this estimator for every sample, how much does the answer vary, on average?"



But there's a problem here...

Knowing the standard error requires knowing what happens across many separate samples. But we've only got our one sample!

**So how can we ever calculate the standard error?**



*Two roads diverged in a yellow wood And sorry I could not travel both And be one traveler, long I stood And looked down one as far as I could To where it bent in the under-growth...*

– Robert Frost, The Road Not Taken, 1916

Quantifying our uncertainty would seem to require knowing all the roads not taken—an impossible task.

# The bootstrap



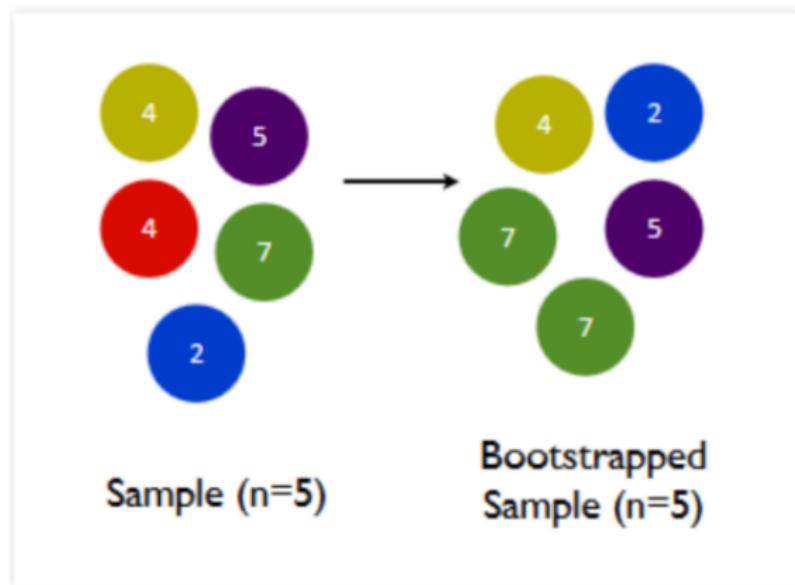
**Problem:** we can't take repeated samples of size  $N$  from the population, to see how our estimate changes across samples.

**Seemingly hacky solution:** Take repeated samples of size  $N$ , with replacement, from the sample itself, and see how our estimate changes across samples. This is something we can easily simulate on a computer (with R).

Basically, we pretend that our sample is the whole population and we charge ahead! This is called bootstrap resampling, or just bootstrapping.

# Sampling with replacement is key!

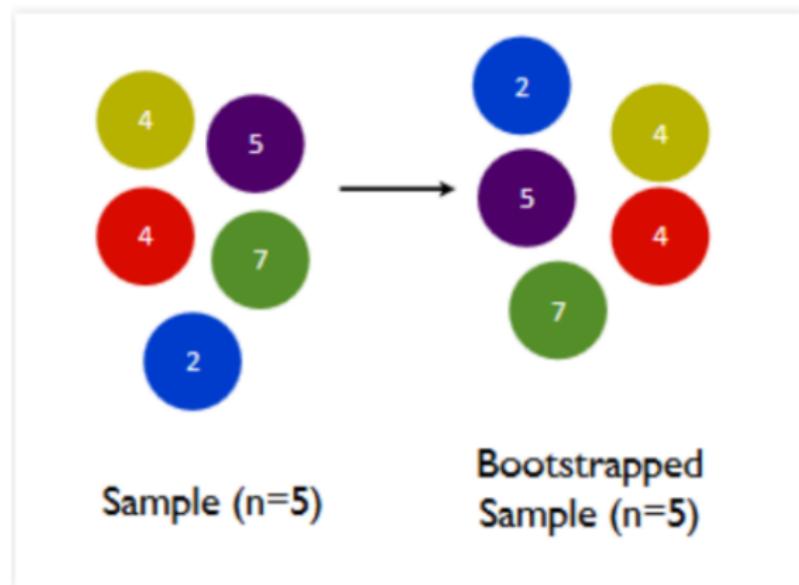
bootstrapped sample 1



# Sampling with replacement is key!

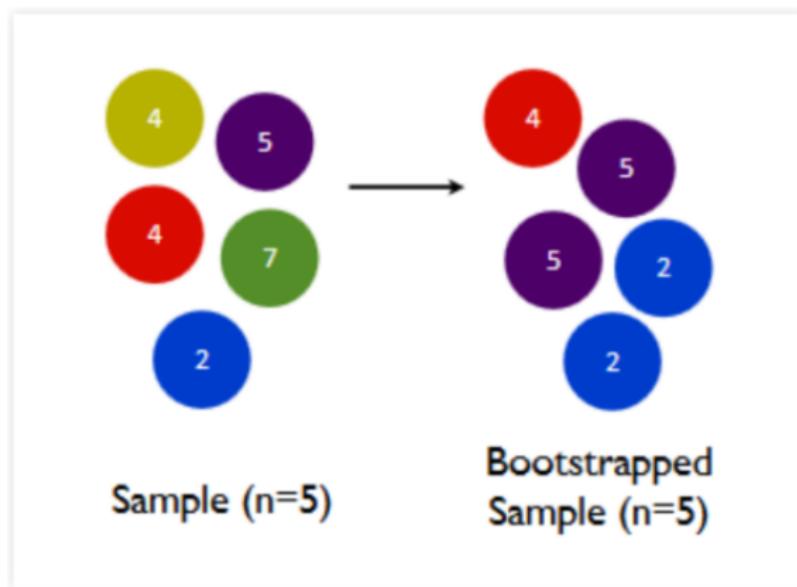


bootstrapped sample 2

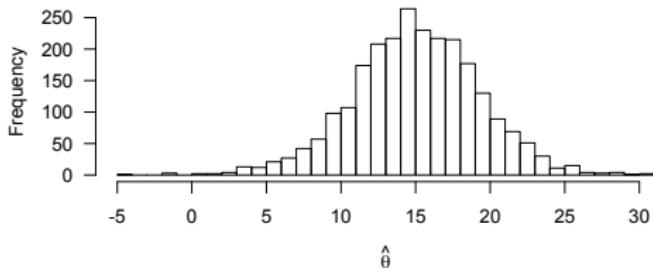
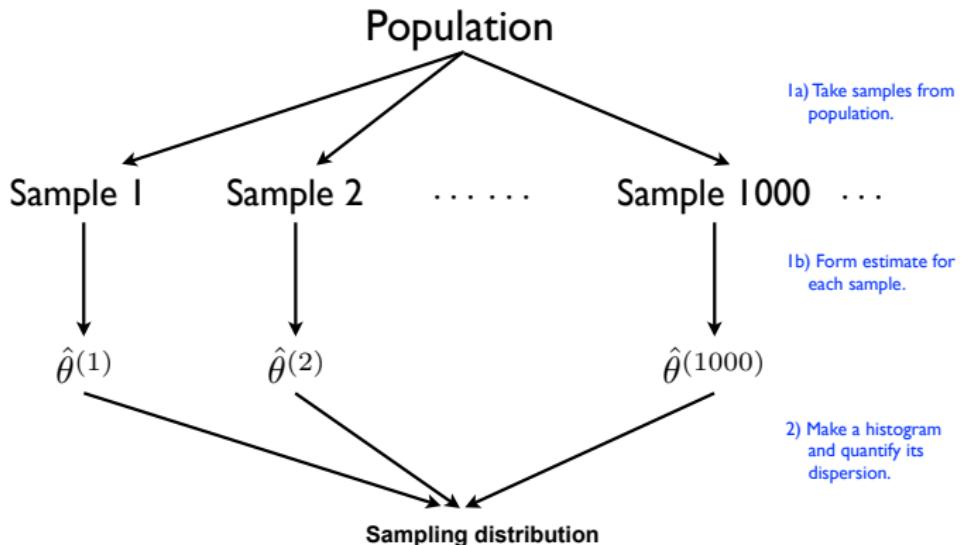


# Sampling with replacement is key!

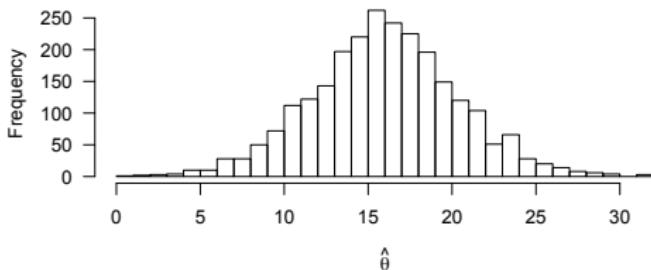
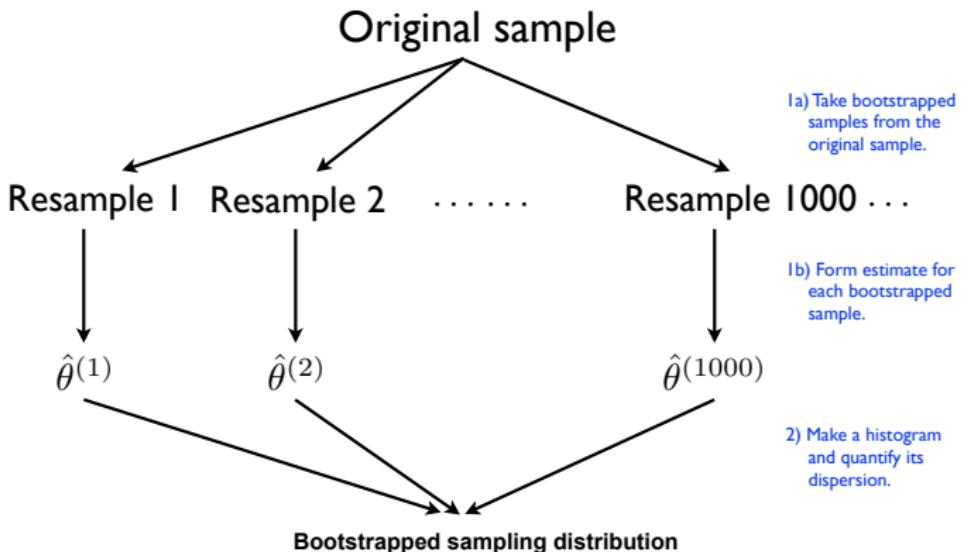
bootstrapped sample 3



# The true sampling distribution



# The bootstrapped sampling distribution



# The bootstrapped sampling distribution



- Each bootstrapped sample has its own pattern of duplicates and omissions from the original sample.
- These duplicates and omissions create variability in  $\hat{\theta}$  from one bootstrapped sample to the next.
- This variability mimics the true sampling variability you'd expect to see across real repeated samples from the population.

Let's check out `bootstrap.R` to see this in action!



## Multiple linear regression

# The multiple linear regression model



Many problems involve more than one independent variable or factor which affects the dependent or response variable.

- More than size to predict house price!
- Demand for a product given prices of competing brands, advertising, household attributes, etc.

In SLR, the conditional mean of  $Y$  depends on  $X$ . The Multiple Linear Regression (MLR) model extends this idea to include more than one independent variable.

# The MLR model



Same as always, but with more covariates.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Recall the key assumptions of our linear regression model:

- The conditional mean of  $Y$  is **linear** in the  $X_j$  variables.
- The errors (deviations from line)
  - are normally distributed
  - independent from each other
  - identically distributed (i.e., they have constant variance)

$$Y|(X_1 \dots X_p) \sim N(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p, \sigma^2)$$



Our interpretation of regression coefficients can be extended from the simple single covariate regression case:

$$\beta_j = \frac{\partial E[Y|X_1, \dots, X_p]}{\partial X_j}$$

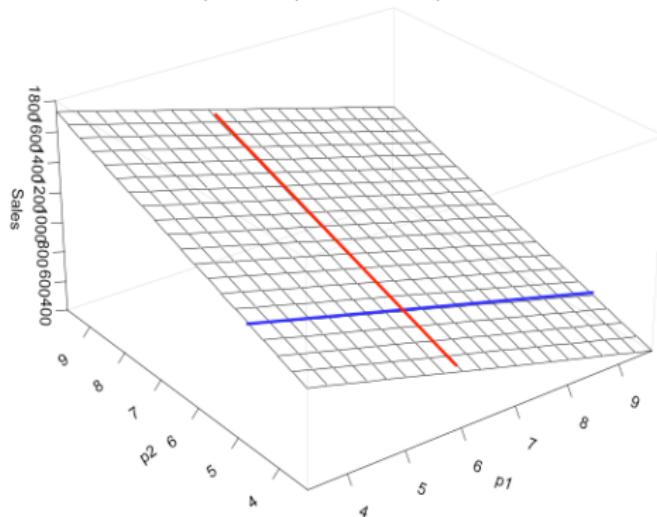
Holding all other variables constant,  $\beta_j$  is the average change in  $Y$  per unit change in  $X_j$ .

# The MLR model

If  $p = 2$ , we can plot the regression surface in 3D.

Consider sales of a product as predicted by price of this product ( $P_1$ ) and the price of a competing product ( $P_2$ ).

$$\text{Sales} = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \epsilon$$



# Least squares again!

$$Y = \beta_0 + \beta_1 X_1 \dots + \beta_p X_p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

## How do we estimate the MLR model parameters?

The principle of **least squares** is exactly the same as before:

- Define the fitted values
- Find the best fitting plane by minimizing the sum of squared residuals

# Least squares again!

The data...

p1	p2	Sales
5.1356702	5.2041860	144.48788
3.4954600	8.0597324	637.24524
7.2753406	11.6759787	620.78693
4.6628156	8.3644209	549.00714
3.5845370	2.1502922	20.42542
5.1679168	10.1530371	713.00665
3.3840914	4.9465690	346.70679
4.2930636	7.7605691	595.77625
4.3690944	7.4288974	457.64694
7.2266002	10.7113247	591.45483

... ... ...

The model:  $\text{Sales}_i = \beta_0 + \beta_1 P1_i + \beta_2 P2_i + \epsilon_i$ ,  $\epsilon \sim N(0, \sigma^2)$

# Fitting the MLR model

```
data = read.csv('PricesSales.csv')
fit = lm(Sales~p1+p2,data)
summary(fit)

##
## Call:
## lm(formula = Sales ~ p1 + p2, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -66.916 -15.663 -0.509  18.904  63.302 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 115.717    8.548   13.54 <2e-16 ***
## p1          -97.657    2.669  -36.59 <2e-16 ***
## p2           108.800   1.409   77.20 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 28.42 on 97 degrees of freedom
## Multiple R-squared:  0.9871, Adjusted R-squared:  0.9869 
## F-statistic: 3717 on 2 and 97 DF,  p-value: < 2.2e-16
```

# Plug-in Prediction in MLR



Suppose that by using advanced corporate espionage tactics, I discover that my competitor will charge \$10 the next quarter. After some marketing analysis I decide to charge \$8. **How much will I sell?**

Our model is:

$$\text{Sales} = \beta_0 + \beta_1 P1 + \beta_2 P2 + \epsilon$$

with  $\epsilon \sim N(0, \sigma^2)$

Our estimates are  $b_0 = 115$ ,  $b_1 = -97$ ,  $b_2 = 109$  and  $s = 28$  which leads to

$$\text{Sales} = 115 + -97 * P1 + 109 * P2 + \epsilon$$

with  $\epsilon \sim N(0, 28^2)$

# Plug-in Prediction in MLR



By plugging-in the numbers,

$$\begin{aligned} Sales &= 115 + -97 * 8 + 109 * 10 + \epsilon \\ &= 437 + \epsilon \end{aligned}$$

$$Sales | (P1 = 8, P2 = 10) \sim N(437, 28^2)$$

and the 95% Prediction Interval is  $(437 \pm 2 * 28)$

$$381 < Sales < 493$$

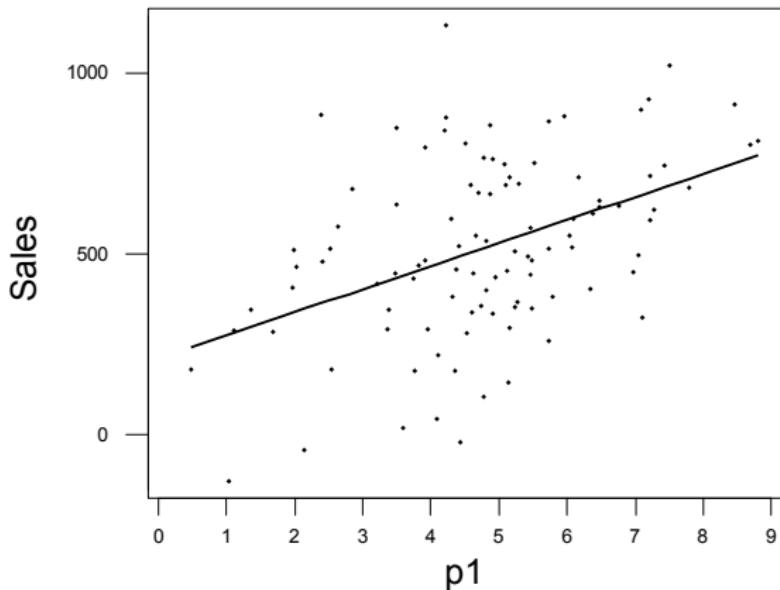
## But what about the right-hand-side?

It is also important to understand and interpret the coefficients, i.e., what is happening on the “right-hand-side” of our model ...

- **Sales** : units sold in excess of a baseline
- **P1**: our price in \$ (in excess of a baseline price)
- **P2**: competitors price (again, over a baseline)

But what about the right-hand-side?

If we regress Sales on our own price, we obtain a somewhat surprising conclusion... **the higher the price the more we sell!**



→ It looks like we should just raise our prices, right?

# Understanding multiple regression



The regression equation for Sales on own price (P1) is:

$$Sales = 211 + 63.7P1$$

If now we add the competitors price to the regression we get

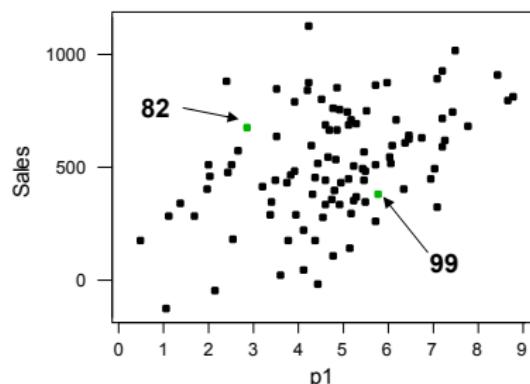
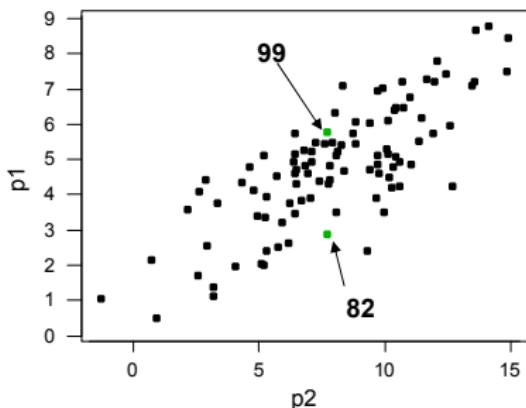
$$Sales = 116 - 97.7P1 + 109P2$$

Does this look better? How did it happen? Remember:  $-97.7$  is the affect on sales of a change in  $P1$  **with  $P2$  held fixed!**

# Understanding multiple regression

How can we see what is going on? Let's compare Sales in two different observations: weeks 82 and 99.

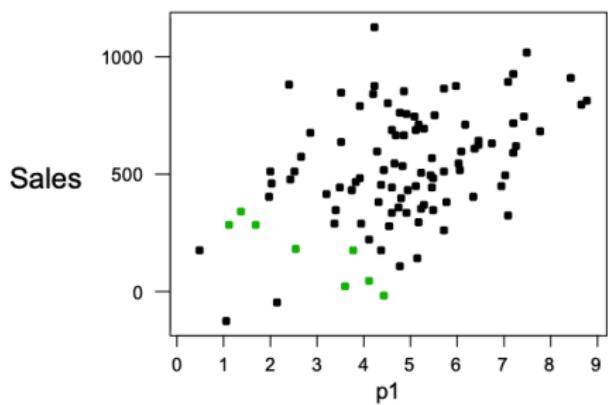
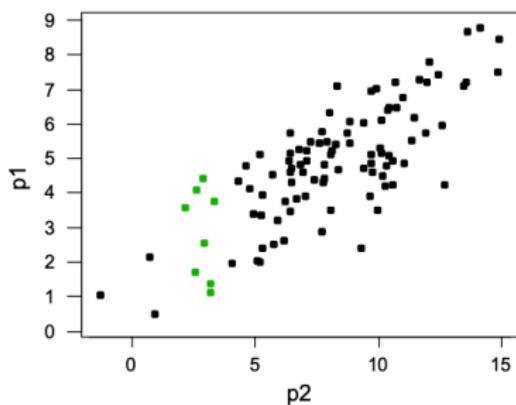
We see that an **increase** in  $P1$ , holding  $P2$  **constant**, corresponds to a drop in Sales!



Note the strong relationship (dependence) between  $P1$  and  $P2$ !

# Understanding multiple regression

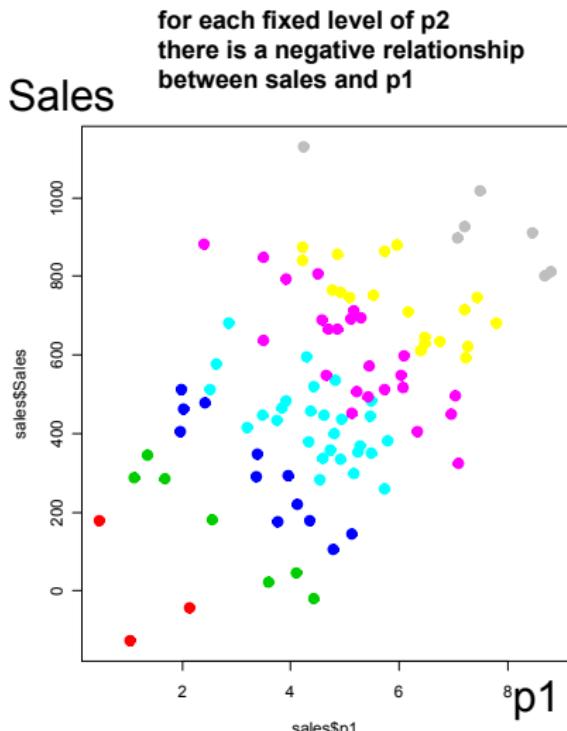
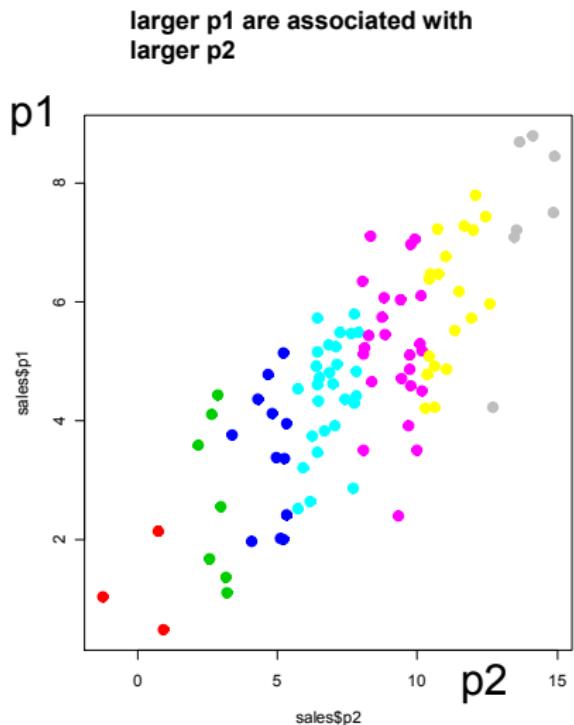
Let's look at a subset of points where  $P1$  varies and  $P2$  is held approximately constant...



For a fixed level of  $P2$ , variation in  $P1$  is negatively correlated with Sales!

# Understanding multiple regression

Below, different colors indicate different ranges for  $P2$ ...



# Understanding multiple regression



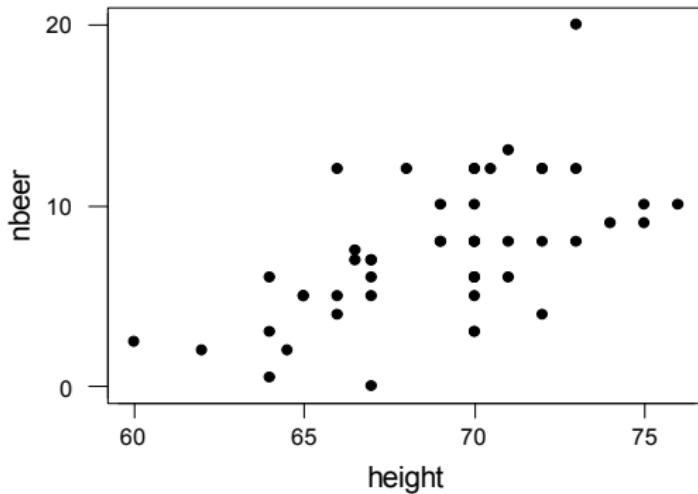
## Summary:

- A larger  $P_1$  is associated with larger  $P_2$  and the overall effect leads to bigger sales
- With  $P_2$  held fixed, a larger  $P_1$  leads to lower sales
- MLR does the trick and unveils the **correct** economic relationship between Sales and prices!

# Example: Beers, height, weight, and getting drunk

Beer data (from a graduate school class)

- **nbeer** – number of beers before getting drunk
- **height** and **weight**



# R output: Yes!



```
data = read.csv('nbeer.csv')
fit = lm(nbeer~height,data)
summary(fit)

##
## Call:
## lm(formula = nbeer ~ height, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.164 -2.005 -0.093  1.738  9.978 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -36.9200    8.9560  -4.122 0.000148 ***
## height       0.6430    0.1296   4.960 9.23e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.109 on 48 degrees of freedom
## Multiple R-squared:  0.3389, Adjusted R-squared:  0.3251 
## F-statistic: 24.6 on 1 and 48 DF,  p-value: 9.23e-06
```

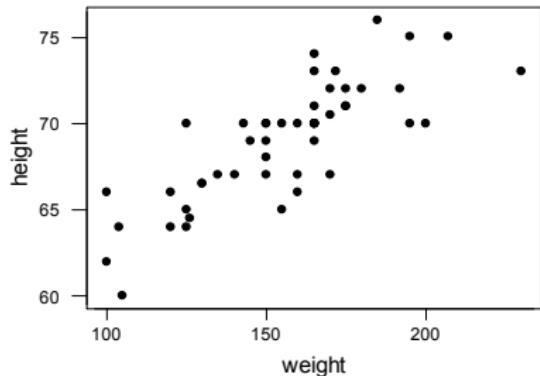
## R output: What about now?



```
data = read.csv('nbeer.csv')
fit = lm(nbeer~height+weight,data)
summary(fit)

##
## Call:
## lm(formula = nbeer ~ height + weight, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.5080 -2.0269  0.0652  1.5576  5.9087
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.18709  10.76821 -1.039 0.304167
## height       0.07751   0.19598  0.396 0.694254
## weight       0.08530   0.02381  3.582 0.000806 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.784 on 47 degrees of freedom
## Multiple R-squared:  0.4807, Adjusted R-squared:  0.4586
## F-statistic: 21.75 on 2 and 47 DF,  p-value: 2.056e-07
```

# Understanding multiple regression



The correlations:

	nbeer	weight
weight	0.692	
height	0.582	

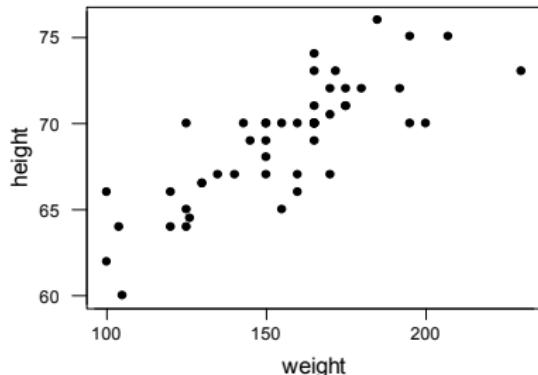
0.806  
→

*The two x's are  
highly correlated !!*

If we regress “beers” only on height we see an effect. Taller heights go with more beers.

However, when height goes up weight tends to go up as well... in the first regression, height was a proxy for the real **cause** of drinking ability. Bigger people can drink more and weight is a more accurate measure of “bigness.”

# Understanding multiple regression



The correlations:

	nbeer	weight
weight	0.692	
height	0.582	0.806

*The two x's are  
highly correlated !!*

In the multiple regression, when we consider only the variation in height that is not associated with variation in weight, we see no relationship between height and beers.

## R output: Why is this a better model than height + weight?



```
data = read.csv('nbeer.csv')
fit = lm(nbeer~weight,data)
summary(fit)

##
## Call:
## lm(formula = nbeer ~ weight, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -8.7709 -2.0304 -0.0742  1.6580  5.6556 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -7.02070   2.21329  -3.172  0.00264 ***
## weight       0.09289   0.01399   6.642  2.6e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.76 on 48 degrees of freedom
## Multiple R-squared:  0.4789, Adjusted R-squared:  0.4681 
## F-statistic: 44.12 on 1 and 48 DF,  p-value: 2.602e-08
```

## Summary slide



In general, when we see a relationship between  $y$  and  $x$  (or  $x$ 's), that relationship may be driven by variables “lurking” in the background which are related to your current  $x$ 's.

This makes it hard to reliably find “causal” relationships. Any correlation (association) you find could be caused by other variables in the background... **correlation is NOT causation**

Any time a report says two variables are related and there's a suggestion of a “causal” relationship, ask yourself whether or not other variables might be the real reason for the effect.

Multiple regression allows us to **control** for all important variables by including them into the regression. **“Once we control for weight, height and beers are NOT related”!**

# Correlation is NOT causation



I USED TO THINK  
CORRELATION IMPLIED  
CAUSATION.



THEN I TOOK A  
STATISTICS CLASS.  
NOW I DON'T.



SOUNDS LIKE THE  
CLASS HELPED.

WELL, MAYBE.

