

## Section 3: Simple Linear Regression

Tengyuan Liang, Chicago Booth

<https://tyliang.github.io/BUS41000/>

Suggested Reading:  
OpenIntro Statistics, Chapters 7  
Statistics for Business, Part IV

# 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 prediction questions:

- ▶ For how much will my house sell? (Zillow's Zestimate algorithm)
- ▶ How many runs per game will the Red Sox score this year?
- ▶ Will this person like that movie? (Netflix rating system)

Explanation and understanding:

- ▶ What is the impact of MBA on income?
- ▶ Is Bitcoin a hedge against market downturns? (Crypto portfolio strategy)
- ▶ Does Walmart discriminate different groups regarding salaries?

# 1st Example: Predicting House Prices

## Problem:

- ▶ Predict market price based on observed characteristics
- ▶ Real-world: Zillow's "Zestimate" algorithm

## 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

## 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 labelled:

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

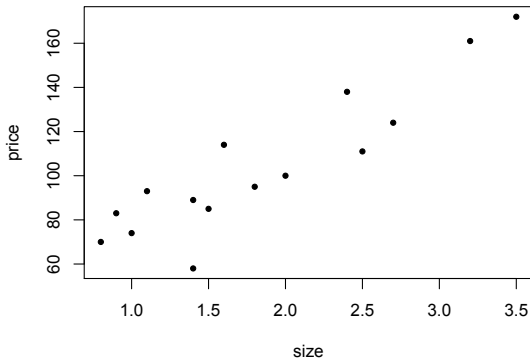
# Predicting House Prices

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

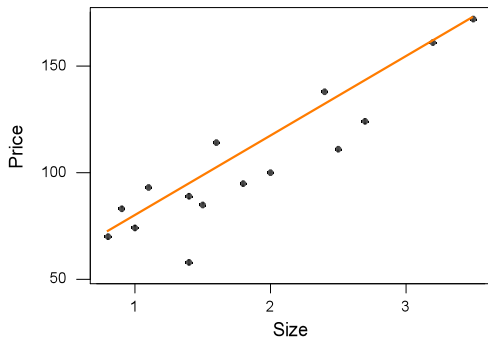
$X$  = explanatory or input variables

A linear relationship is written

$$Y = b_0 + b_1X$$

# Linear Prediction

Appears to be a linear relationship between price and size:  
As size goes up, price goes up.



The line shown was fit by the “eyeball” method.

# Linear Prediction

Recall that the equation of a line is:

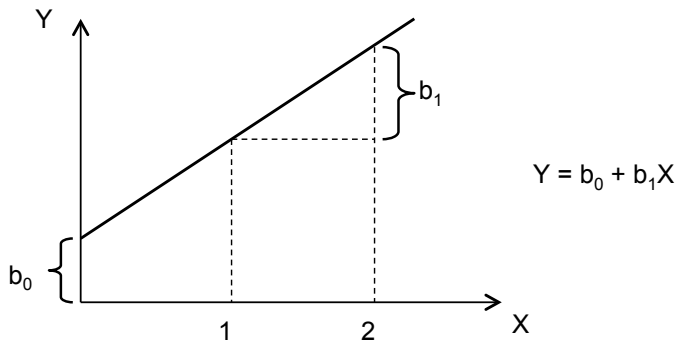
$$Y = b_0 + b_1X$$

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



Our "eyeball" line has  $b_0 = 35$ ,  $b_1 = 40$ .

## Linear Prediction

We can now predict the price of a house when we know only the size; just read the value off the line that we've drawn.

For example, given a house with of size  $X = 2.2$ .

Predicted price  $\hat{Y} = 35 + 40(2.2) = 123$ .

Note: Conversion from 1,000 sq ft to \$1,000 is done for us by the slope coefficient ( $b_1$ )

# 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.

# 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**.

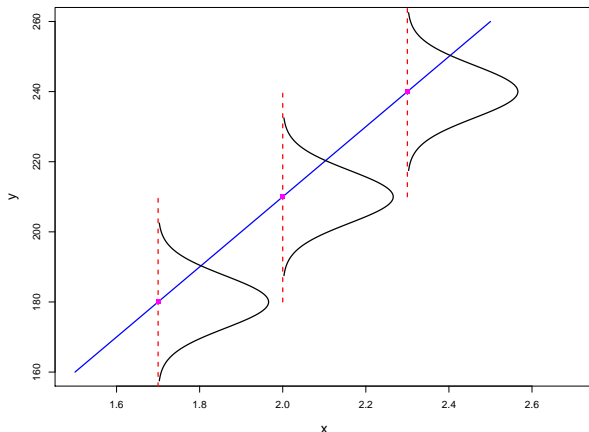
Simple Linear Regression Model:  $Y = \beta_0 + \beta_1 X + \varepsilon$

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

- ▶  $\beta_0 + \beta_1 X$  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 conditional distribution for  $Y$  given  $X$  is Normal:

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

# 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  $87.5 < Y < 127.5$

# Linear Prediction

Can we do better than the eyeball method?

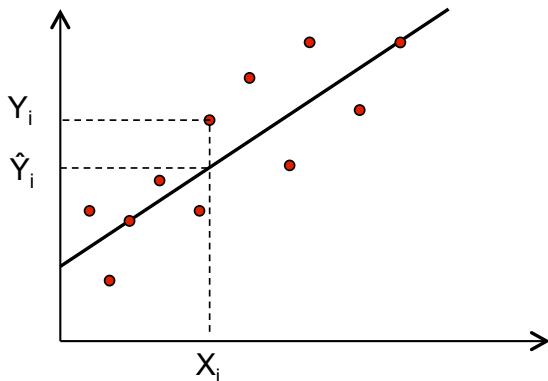
We desire a strategy for estimating the slope and intercept parameters in the model  $\hat{Y} = b_0 + b_1X$

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

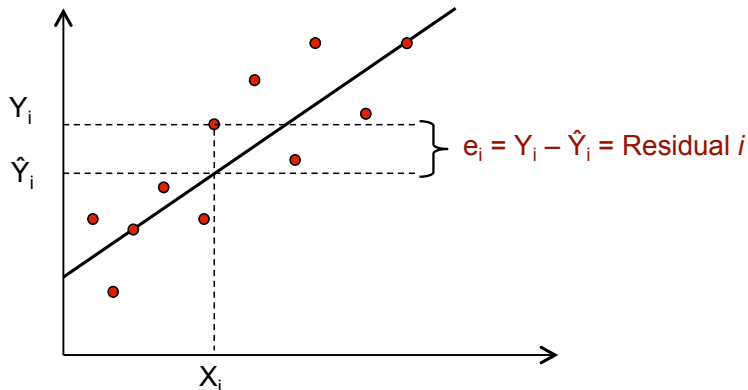
What is the “fitted value”?



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

# 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.

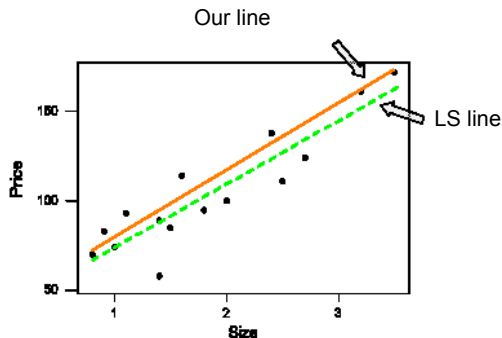
Least Squares chooses  $b_0$  and  $b_1$  to minimize  $\sum_{i=1}^N e_i^2$

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

# Least Squares

LS chooses a different line from ours:

- ▶  $b_0 = 38.88$  and  $b_1 = 35.39$
- ▶ What do  $b_0$  and  $b_1$  mean again?



# Least Squares – Excel Output

## SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.909209967
R Square	0.826662764
Adjusted R Square	0.81332913
Standard Error	14.13839732
Observations	15

## ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	12393.10771	12393.10771	61.99831126	2.65987E-06
Residual	13	2598.625623	199.8942787		
Total	14	14991.73333			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	38.88468274	9.09390389	4.275906499	0.000902712	19.23849785	58.53086763
Size	35.38596255	4.494082942	7.873900638	2.65987E-06	25.67708664	45.09483846

# Real-World Case: Zillow Offers Failure (2021)

## Background:

- ▶ Zillow launched “Zillow Offers” (iBuying) in April 2018
- ▶ Used Zestimate algorithm to directly buy homes, renovate, and resell
- ▶ Goal: Profit from algorithmic house price predictions

## What Happened?

- ▶ **November 2021:** Zillow shut down Zillow Offers
- ▶ Lost **\$420 million in Q3 2021 alone**
- ▶ Laid off **25% of employees** (~2,500 workers)
- ▶ Owned ~7,000 homes that needed to be sold at losses

# Zillow's Algorithmic Failure: What Went Wrong?

## The Prediction Problem:

- ▶ Zestimate model:  $\widehat{Price} = f(size, beds, location, \dots)$
- ▶ Worked reasonably well for *estimating* ( $R^2 \approx 0.80$ )
- ▶ But prediction intervals were too wide for profitable trading

## Key Issues:

1. Ignored prediction uncertainty:  $s = \$14K$  means  $\pm \$28K$  range (95% CI)
2. Market timing risk: Model couldn't predict rapid price changes (2020-2021)
3. Operational costs: Renovation costs exceeded predictions
4. Volume problem: Bought ~27,000 homes at scale

# Zillow: Statistical vs. Business Success

## The Statistical Model Was "Good":

- ▶ High  $R^2$  (explained variation)
- ▶ Statistically significant coefficients
- ▶ Worked well for information/estimation

## But Business Model Failed:

- ▶ \$420M loss = 27,000 homes  $\times$   $\sim$ \$15K loss per home
- ▶ Prediction intervals matter more than point estimates!
- ▶ Statistical accuracy  $\neq$  profitable trading strategy
- ▶ Scale amplified small errors:  $\pm$ \$15K error  $\times$  27,000 homes = \$405M risk

**Lesson:** Always report and act on prediction intervals, not just  $\hat{Y}$

## 2nd Example: Offensive Performance in Baseball

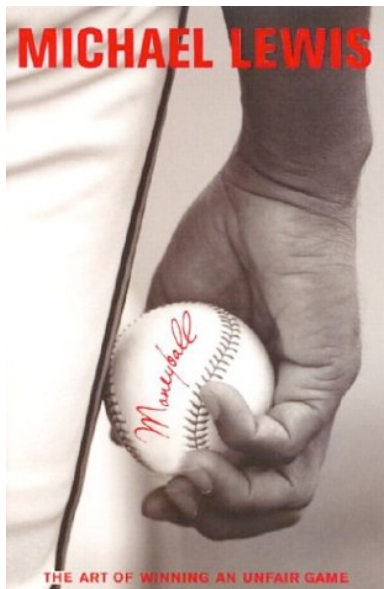
### 1. Problems:

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

### 2. Solutions:

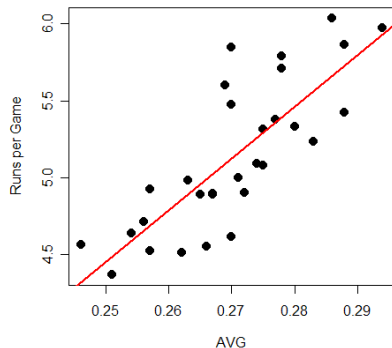
- ▶ Compare *prediction rules* that forecast runs as a function of either AVG (batting average), SLG (slugging percentage) or OBP (on base percentage)

## 2nd Example: Offensive Performance in Baseball



# Baseball Data – Using AVG

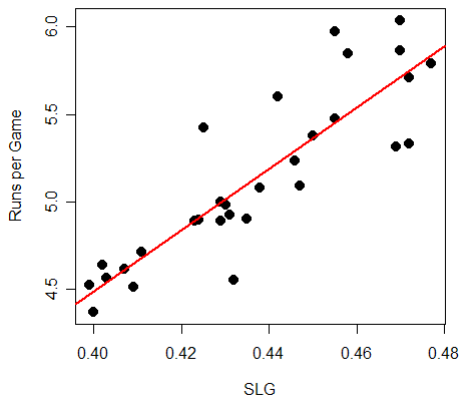
Each observation corresponds to a team in MLB. Each quantity is the average over a season.



►  $Y = \text{runs per game}; X = \text{AVG (average)}$

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

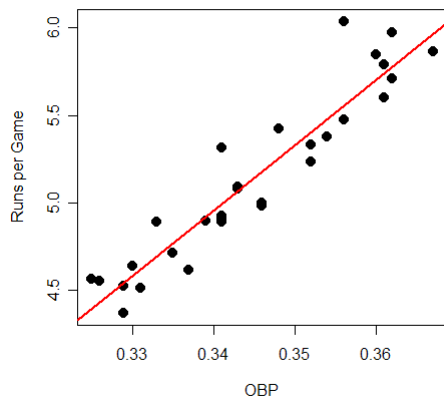
# Baseball Data – Using SLG



- ▶  $Y$  = runs per game
- ▶  $X$  = SLG (slugging percentage)

LS fit:  $\text{Runs/Game} = -2.52 + 17.54 \text{ SLG}$

# Baseball Data – Using OBP



- ▶  $Y$  = runs per game
- ▶  $X$  = OBP (on base percentage)

LS fit:  $\text{Runs/Game} = -7.78 + 37.46 \text{ OBP}$

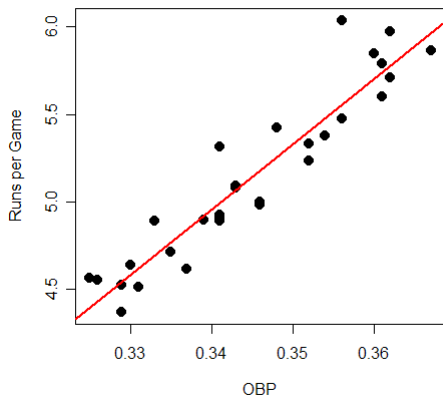
- ▶ What is the best prediction rule?
- ▶ Let's compare the predictive ability of each model using the average squared error

$$\frac{1}{N} \sum_{i=1}^N e_i^2 = \frac{\sum_{i=1}^N \left( \widehat{Runs_i} - Runs_i \right)^2}{N}$$

## Place your Money on OBP!!!

Average Squared Error	
AVG	0.083
SLG	0.055
OBP	0.026

# Linear Prediction



$$\hat{Y}_i = b_0 + b_1 X_i$$

- ▶  $b_0$  is the intercept and  $b_1$  is the slope
- ▶ We find  $b_0$  and  $b_1$  using *Least Squares*

# The Least Squares Criterion

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

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

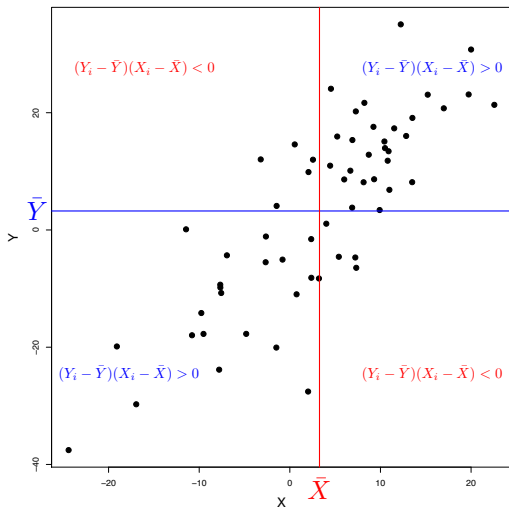
where,

- ▶  $\bar{X}$  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  $Y$

# 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?

# Correlation

Correlation is the standardized covariance:

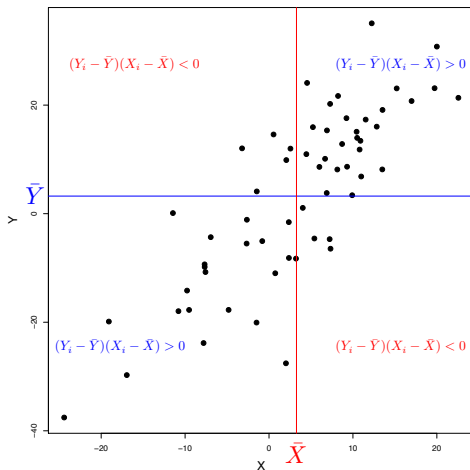
$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{s_x^2 s_y^2}} = \frac{\text{cov}(X, Y)}{s_x s_y}$$

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

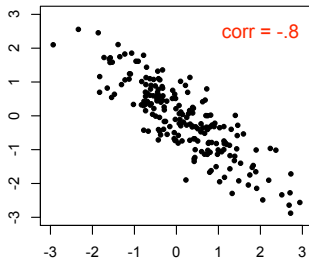
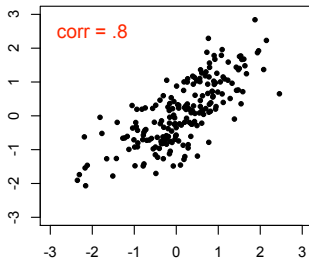
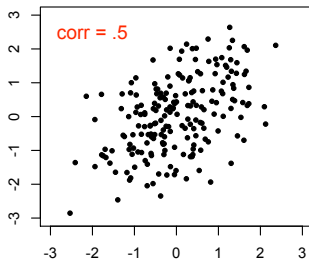
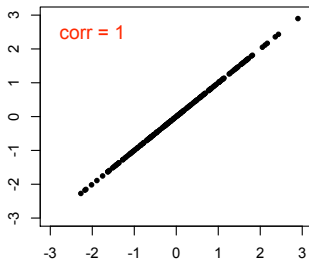
This gives the direction (- or +) and strength (0 → 1) of the linear relationship between  $X$  and  $Y$ .

# Correlation

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



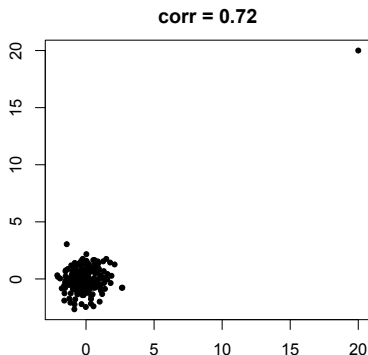
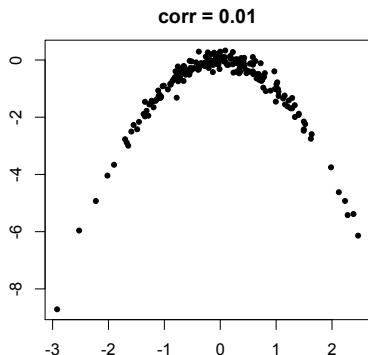
# Correlation



# Correlation

Only measures **linear** relationships:

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



Also be careful with influential observations.

**Excel:** correl, stdev, ...

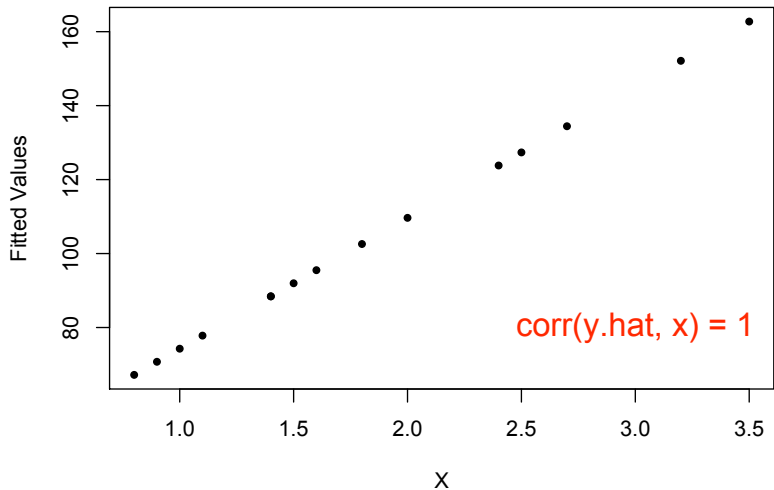
**R:** cor, sd, ...

## More on Least Squares

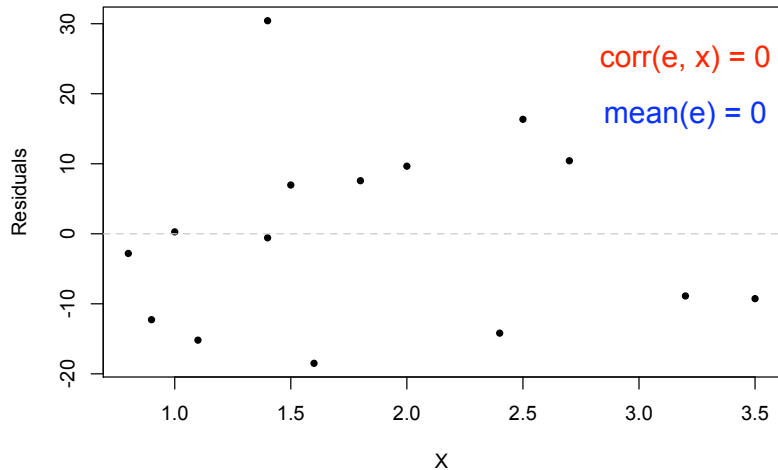
From now on, terms "fitted values" ( $\hat{Y}_i$ ) and "residuals" ( $e_i$ ) refer to those obtained from the least squares line.

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

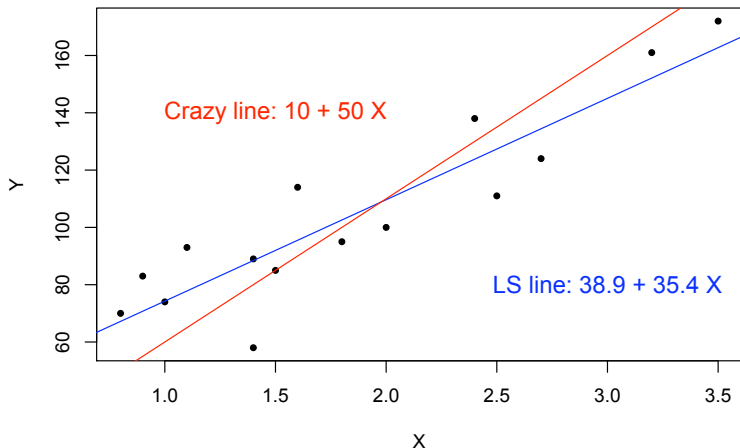


# The Residuals and X



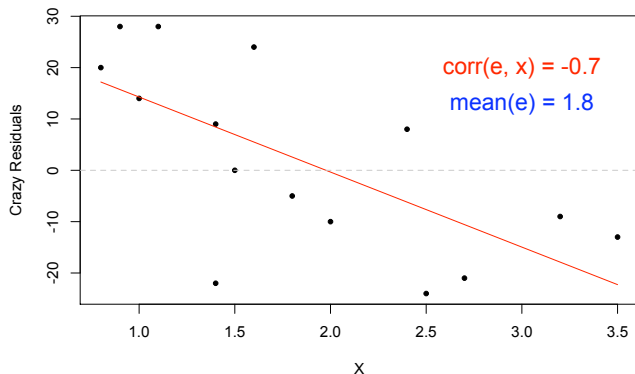
# Why?

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

## Summary: LS is the best we can do!!

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

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.  $\text{corr}(\hat{Y}, e) = 0$ ,

$$\begin{aligned}\text{var}(Y) &= \text{var}(\hat{Y} + e) = \text{var}(\hat{Y}) + \text{var}(e) \\ \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1} &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2}{n-1} + \frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n-1}\end{aligned}$$

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

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{\substack{\text{Total Sum of} \\ \text{Squares} \\ \text{SST}}} = \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{\substack{\text{Regression SS} \\ \text{SSR}}} + \underbrace{\sum_{i=1}^n e_i^2}_{\substack{\text{Error SS} \\ \text{SSE}}}$$

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:

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

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

# Back to the House Data

## SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.909209967
R Square	0.826662764
Adjusted R Square	0.81332913
Standard Error	14.13839732
Observations	15

## ANOVA

	df	SS	MS	F	Significance F
Regression	1	12393.10771	12393.10771	61.99831126	2.65987E-06
Residual	13	2598.625623	199.8942787		
Total	14	14991.73333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	38.86468274	9.09390389	4.275906499	0.000902712	19.23849785	58.53086763	19.23849785	58.53086763
X Variable 1	35.38596255	4.494082942	7.873900638	2.65987E-06	25.67708664	45.09483846	25.67708664	45.09483846

SSR

SST

SSE

$$R^2 = \frac{SSR}{SST} = 0.82 = \frac{12395}{14991}$$

## 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

# 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).

Assumptions:

- ▶ **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  $\varepsilon_i$

# 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  $\sigma^2$  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.

# Estimation of Error Variance

Where is  $s$  in the Excel output?

## SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.909209967
R Square	0.826662764
Adjusted R Square	0.81332913
Standard Error	14.13839732
Observations	15

$s$

## ANOVA

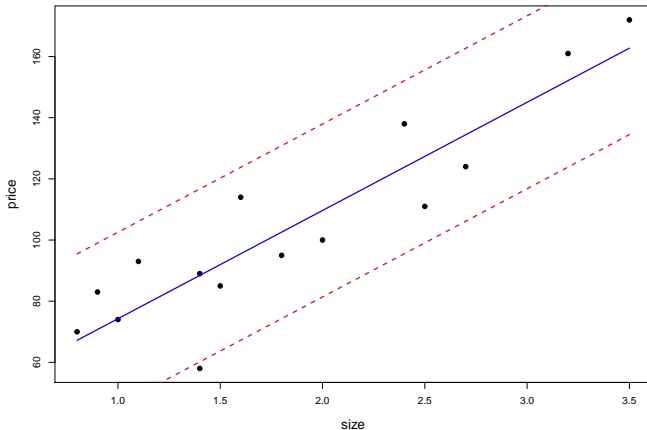
	df	SS	MS	F	Significance F
Regression	1	12393.10771	12393.10771	61.99831126	2.65987E-06
Residual	13	2598.625623	199.8942787		
Total	14	14991.73333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	38.88468274	9.09390389	4.275906499	0.000902712	19.23849785	58.53086763	19.23849785	58.53086763
X Variable 1	35.38596255	4.494082942	7.873900638	2.65987E-06	25.67708664	45.09483846	25.67708664	45.09483846

Remember that whenever you see “standard error” read it as estimated standard deviation:  $\sigma$  is the standard deviation.

# One Picture Summary of SLR

- ▶ The plot below has the house data, the fitted regression line ( $b_0 + b_1X$ ) and  $\pm 2 * s \dots$
- ▶ From this picture, what can you tell me about  $\beta_0$ ,  $\beta_1$  and  $\sigma^2$ ?  
How about  $b_0$ ,  $b_1$  and  $s^2$ ?



# Sampling Distribution of Least Squares Estimates

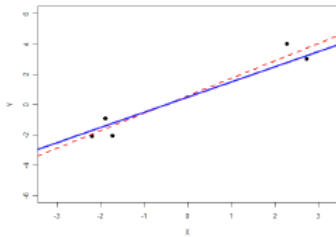
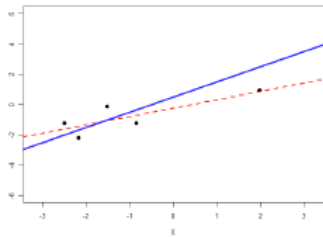
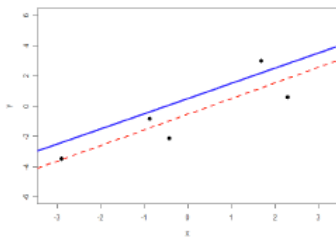
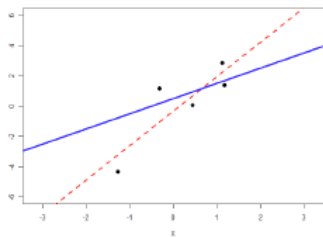
How much do our estimates depend on the particular random sample that we happen to observe? Imagine:

- ▶ 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

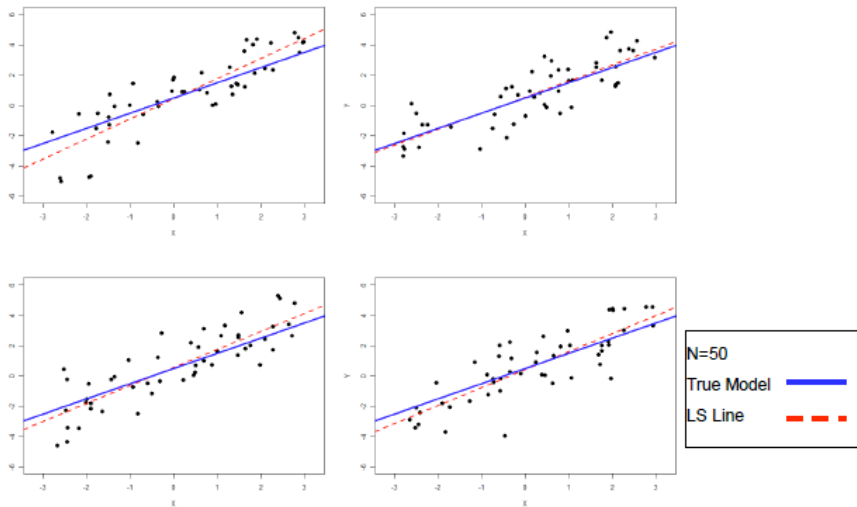


N=5

True Model

LS Line

# Sampling Distribution of Least Squares Estimates



# Sampling Distribution of $b_1$

The sampling distribution of  $b_1$  describes how estimator  $b_1 = \hat{\beta}_1$  varies over different samples with the  $X$  values fixed.

It turns out that  $b_1$  is normally distributed (approximately):  
 $b_1 \sim N(\beta_1, s_{b_1}^2)$ .

- ▶  $b_1$  is unbiased:  $E[b_1] = \beta_1$ .
- ▶  $s_{b_1}$  is the **standard error of  $b_1$** . In general, the standard error is the standard deviation of an estimate. It determines **how close**  $b_1$  is to  $\beta_1$ .
- ▶ This is a number directly available from the regression output.

# Sampling Distribution of $b_1$

Can we intuit what should be in the formula for  $s_{b_1}$ ?

- ▶ How should  $s$  figure in the formula?
- ▶ What about  $n$ ?
- ▶ Anything else?

$$s_{b_1}^2 = \frac{s^2}{\sum (X_i - \bar{X})^2} = \frac{s^2}{(n-1)s_x^2}$$

Three Factors:

sample size ( $n$ ), error variance ( $s^2$ ), and  $X$ -spread ( $s_x$ ).

## Sampling Distribution of $b_0$

The intercept is also **normal** and **unbiased**:  $b_0 \sim N(\beta_0, s_{b_0}^2)$ .

$$s_{b_0}^2 = \text{var}(b_0) = s^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2} \right)$$

What is the intuition here?

# Confidence Intervals

Since  $b_1 \sim N(\beta_1, s_{b_1}^2)$ , Thus:

- ▶ 68% Confidence Interval:  $b_1 \pm 1 \times s_{b_1}$
- ▶ 95% Confidence Interval:  $b_1 \pm 2 \times s_{b_1}$
- ▶ 99% Confidence Interval:  $b_1 \pm 3 \times s_{b_1}$

Same thing for  $b_0$

- ▶ 95% Confidence Interval:  $b_0 \pm 2 \times s_{b_0}$

The confidence interval provides you with a set of plausible values  
for the parameters

# Example: Runs per Game and AVG

## Regression with all points

### SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.798496529
R Square	0.637596707
Adjusted R Square	0.624653732
Standard Error	0.298493066
Observations	30

### ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>significance F</i>
Regression	1	4.38915033	4.38915	49.26199	1.239E-07
Residual	28	2.494747094	0.089098		
Total	29	6.883897424			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-3.936410446	1.294049995	-3.04193	0.005063	-6.587152	-1.2856692
AVG	33.57186945	4.783211061	7.018689	1.24E-07	23.773906	43.369833

$$[b_1 - 2 \times s_{b_1}; b_1 + 2 \times s_{b_1}] \approx [23.77; 43.36]$$

# Testing

Suppose we want to assess whether or not  $\beta_1$  equals a proposed value  $\beta_1^0$ . This is called **hypothesis testing**.

Formally we test the null hypothesis:

$$H_0 : \beta_1 = \beta_1^0$$

vs. the alternative

$$H_1 : \beta_1 \neq \beta_1^0$$

# Testing

That are 2 ways we can think about testing:

1. Building a test statistic... the **t-stat**,

$$t = \frac{b_1 - \beta_1^0}{s_{b_1}}$$

This quantity measures how many standard deviations the estimate ( $b_1$ ) from the proposed value ( $\beta_1^0$ ).

If the absolute value of  $t$  is greater than 2, we need to worry (why?)... we **reject** the hypothesis.

2. Looking at the **confidence interval**. If the proposed value is outside the confidence interval you **reject** the hypothesis.

Notice that this is equivalent to the t-stat. An absolute value for  $t$  greater than 2 implies that the proposed value is outside the confidence interval. . . therefore reject.

This is my preferred approach for the testing problem. You can't go wrong by using the confidence interval!

## Example: Bitcoin vs. S&P 500 (Real Data)

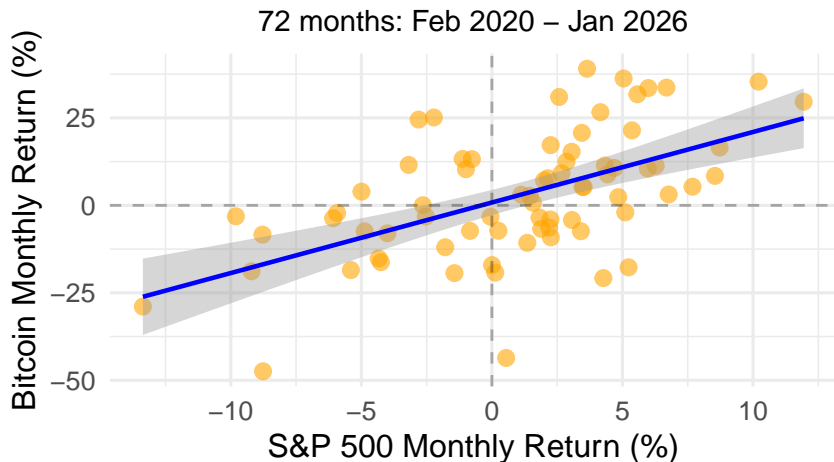
Let's investigate whether Bitcoin is a “hedge” against stock market risk. . .

### The "Digital Gold" Hypothesis:

- ▶ Bitcoin advocates claim it's uncorrelated with stocks
- ▶ Should provide diversification benefits
- ▶ \$1.7 Trillion market as of 2024

**Question:** Does the data support this claim?

## Bitcoin vs. S&P 500: Monthly Returns (2020-2026)



Source: Yahoo Finance (BTC-USD, ^GSPC)

# Bitcoin CAPM Regression

Consider a CAPM regression for Bitcoin:

$$r_{BTC} = \beta_0 + \beta_1 r_{SP500} + \epsilon$$

Let's first test  $\beta_1 = 0$

$H_0 : \beta_1 = 0$ . Is Bitcoin uncorrelated with the market?

$H_1 : \beta_1 \neq 0$

# Bitcoin Regression Results

```
##
## Call:
## lm(formula = Bitcoin ~ SP500, data = crypto_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -45.541  -9.227  -0.941   10.242   30.890
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.8122     1.8042    0.45   0.654
## SP500          2.0160     0.3587    5.62 3.61e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.96 on 70 degrees of freedom
## Multiple R-squared:  0.3109, Adjusted R-squared:  0.3011
## F-statistic: 31.59 on 1 and 70 DF,  p-value: 3.614e-07
```

# Bitcoin Results: Key Findings

LS fit: Bitcoin Return =  $0.81 + 2.02 \times \text{S\&P500 Return}$

Test  $H_0 : \beta_1 = 0$  (independence)

- ▶  $t = 5.62 \dots$  reject  $\beta_1 = 0!!$
- ▶ 95% confidence interval:  $[1.3; 2.73]$
- ▶ p-value  $< 0.001$

**Conclusion:** Bitcoin **IS** correlated with S&P 500.  
The “uncorrelated with stocks” claim is **FALSE**.

# Bitcoin: Testing Market Risk Equivalence

Now let's test  $\beta_1 = 1$ . What does that mean?

$H_0 : \beta_1 = 1$  Bitcoin has same risk as the market.

$H_1 : \beta_1 \neq 1$  Bitcoin has different risk profile.

►  $t = \frac{b_1 - 1}{s_{b_1}} = \frac{2.02 - 1}{0.36} = 2.83 \dots$  **reject!**

► 95% CI:  $[1.3, 2.73]$  excludes 1

**Conclusion:** Bitcoin  $\beta = 2.02$  means it moves **2.02 times the market**.  
It's a **leveraged bet** on risk-on sentiment, NOT a hedge!

# Bitcoin: Business Implications

## Statistical Findings:

- ▶  $\beta = 2.02$  (amplifies market moves by 102%)
- ▶  $R^2 = 0.31$  (market explains 31% of Bitcoin variation)
- ▶ Correlation = 0.56

## Business Translation:

- ▶ If S&P 500 drops 10%, expect Bitcoin to drop ~20%
- ▶ Portfolio “diversification” with Bitcoin **increases** risk
- ▶ “Digital gold” narrative contradicted by data (2020-2026)
- ▶ Institutional investors using Bitcoin for diversification are **misguided**

**Data Period Note:** Feb 2020 - Jan 2026 (72 months)

Correlation increased after institutional adoption began.

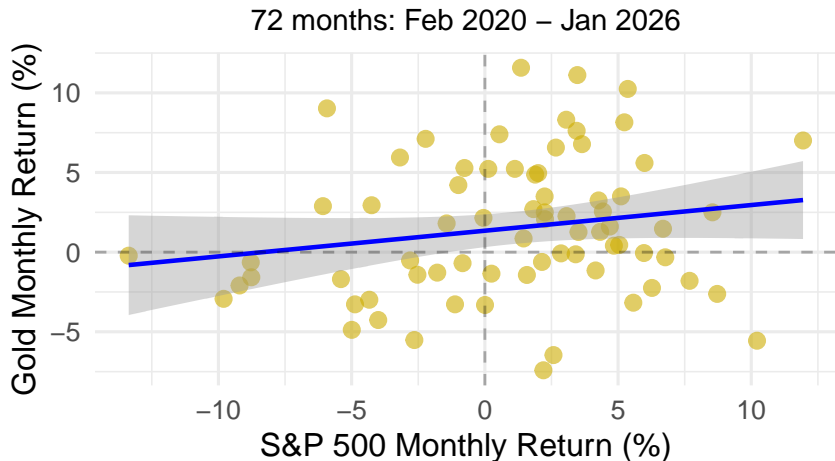
# Testing the Gold Hedge

If Bitcoin is truly “digital gold,” it should behave like actual gold.

Test:  $H_0 : \beta_1 = 0$  (Is gold uncorrelated with S&P 500?)

Data: GLD ETF, Feb 2020 - Jan 2026 (72 months)

## Gold vs. S&P 500: Monthly Returns (2020-2026)



Source: Yahoo Finance (GLD ETF, ^GSPC)

## Gold Regression Results & Findings

```
##
```

```
## Call:
```

```
## lm(formula = Gold ~ SP500, data = gold_data)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
```

```
## -9.1151 -2.7049 -0.6067  3.2354 10.0129
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   1.3448     0.5168   2.602  0.0113 *
```

```
## SP500          0.1612     0.1028   1.569  0.1211
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
```

```
## Residual standard error: 4.286 on 70 degrees of freedom
```

```
## Multiple R-squared:  0.03398,    Adjusted R-squared:  0.02018
```

```
## F-statistic: 2.463 on 1 and 70 DF,  p-value: 0.1211
```

**Result:**  $t = 1.57$ ,  $p = 0.121$

## Bitcoin vs. Gold: The Results

Metric	Bitcoin	Gold
$\beta$	2.02	0.16
<b>t-stat (<math>\beta=0</math>)</b>	5.62***	1.57
<b>p-value</b>	< 0.001	0.121
<b>Correlation</b>	0.56	0.18
<b>R<sup>2</sup></b>	0.31	0.034

**Bitcoin ( $\beta = 2.02$ ):** Moves 2.02 times with the market—amplified leverage on risk.

**Gold ( $\beta = 0.16$ ):** Moves independently—genuine portfolio hedge.

# What The Data Shows

**Bitcoin's claim:** “Digital gold hedge” for portfolio diversification

**Evidence:**  $\beta = 2.02$  with  $p < 0.001$  — it is **NOT** a hedge. It amplifies losses.

**Gold's claim:** “Safe haven” asset independent of stock movement

**Evidence:**  $\beta = 0.16$  with  $p = 0.121$  — it **IS actually uncorrelated**.  
Real hedge.

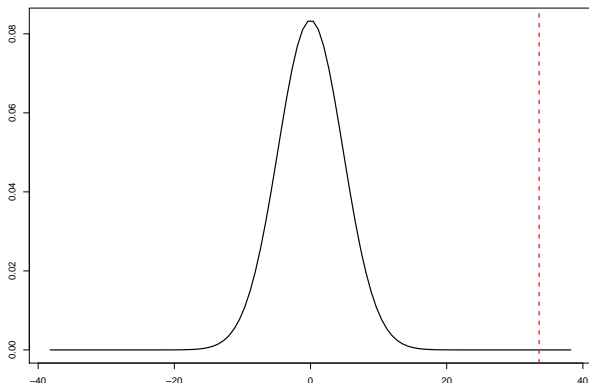
**Portfolio implication:** Use gold ( $\beta \approx 0$ ) for diversification, not Bitcoin ( $\beta = 2$ ).

# P-values

- ▶ The  $p$ -value provides a measure of how weird your estimate is if the null hypothesis is true
- ▶ Small  $p$ -values are evidence against the null hypothesis
- ▶ In the AVG vs. R/G example...  $H_0 : \beta_1 = 0$ . How weird is our estimate of  $b_1 = 33.57$ ?
- ▶ Remember:  $b_1 \sim N(\beta_1, s_{b_1}^2)$ ... If the null was true ( $\beta_1 = 0$ ),  $b_1 \sim N(0, s_{b_1}^2)$

# P-values

- Where is 33.57 in the picture below?



The  $p$ -value is the probability of seeing  $b_1$  equal or greater than 33.57 in absolute terms. Here,  $p\text{-value}=0.000000124!!$

Small  $p$ -value = bad null

# P-values

►  $H_0 : \beta_1 = 0 \dots$  p-value = 1.24E-07... reject!

## SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.798496529
R Square	0.637596707
Adjusted R Square	0.624653732
Standard Error	0.298493066
Observations	30

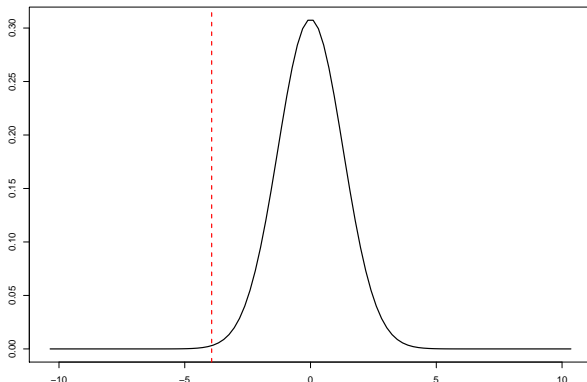
## ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>significance F</i>
Regression	1	4.38915033	4.38915	49.26199	1.239E-07
Residual	28	2.494747094	0.089098		
Total	29	6.883897424			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-3.936410446	1.294049995	-3.04193	0.005063	-6.587152	-1.2856692
AVG	33.57186945	4.783211061	7.018689	1.24E-07	23.773906	43.369833

# P-values

- How about  $H_0 : \beta_0 = 0$ ? How weird is  $b_0 = -3.936$ ?



The  $p$ -value (the probability of seeing  $b_0$  equal or greater than  $-3.936$  in absolute terms) is **0.005**.

Small  $p$ -value = bad null

# P-values

- $H_0 : \beta_0 = 0 \dots$  **p-value = 0.005**... we still reject, but not with the same strength.

## SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.798496529
R Square	0.637596707
Adjusted R Square	0.624653732
Standard Error	0.298493066
Observations	30

## ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>significance F</i>
Regression	1	4.38915033	4.38915	49.26199	1.239E-07
Residual	28	2.494747094	0.089098		
Total	29	6.883897424			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-3.936410446	1.294049995	-3.04193	0.005063	-6.587152	-1.2856692
AVG	33.57186945	4.783211061	7.018689	1.24E-07	23.773906	43.369833

# Testing – Summary

- ▶ Large  $t$  or small  $p$ -value mean the same thing. . .
- ▶  $p$ -value  $< 0.05$  is equivalent to a  $t$ -stat  $> 2$  in absolute value
- ▶ Small  $p$ -value means something weird happen if the null hypothesis was true. . .
- ▶ Bottom line, small  $p$ -value  $\rightarrow$  REJECT! Large  $t \rightarrow$  REJECT!
- ▶ But remember, always look at the confidence interval!

# Forecasting

The **conditional forecasting problem**: Given covariate  $X_f$  and sample data  $\{X_i, Y_i\}_{i=1}^n$ , predict the “future” observation  $y_f$ .

The solution is to use our LS fitted value:  $\hat{Y}_f = b_0 + b_1 X_f$ .

This is the easy bit. The hard (**and very important!**) part of forecasting is assessing uncertainty about our predictions.

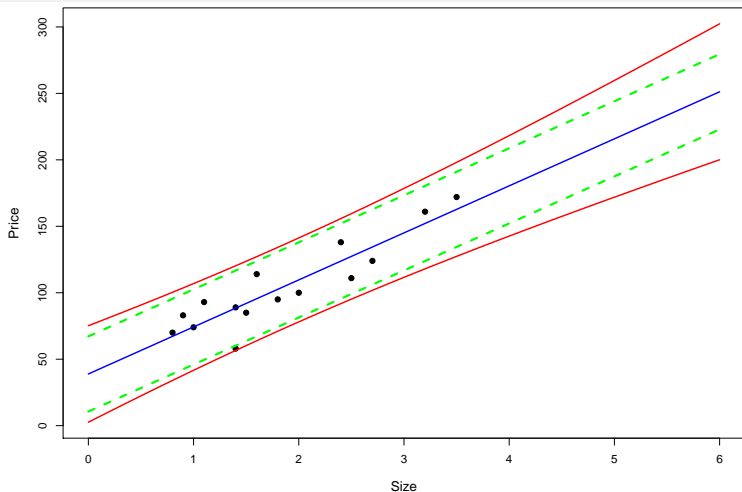
# Forecasting

Just remember that you are uncertain about  $b_0$  and  $b_1$ ! As a practical matter if the confidence intervals are big you should be careful!! Some statistical software will give you a larger (and correct) predictive interval.

A large predictive error variance (high uncertainty) comes from

- ▶ Large  $s$  (i.e., large  $\varepsilon$ 's).
- ▶ Small  $n$  (not enough data).
- ▶ Small  $s_x$  (not enough observed spread in covariates).
- ▶ Large difference between  $X_f$  and  $\bar{X}$ .

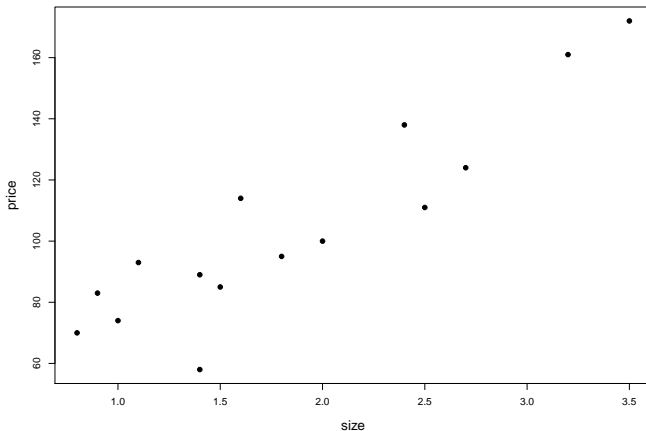
# Forecasting



- ▶ Red lines: prediction intervals
- ▶ Green lines: “plug-in” prediction intervals

# House Data – one more time!

- ▶  $R^2 = 82\%$
- ▶ Great  $R^2$ , we are happy using this model to predict house prices, right?



# House Data – one more time!

- ▶ But,  $s = 14$  leading to a predictive interval width of about US\$60,000!! How do you feel about the model now?
- ▶ As a practical matter,  $s$  is a much more relevant quantity than  $R^2$ . Once again, *intervals* are your friend!

