STA 371G: Statistics and Modeling

Simple Linear Regression: Least Squares Estimation

Mingyuan Zhou
McCombs School of Business
The University of Texas at Austin

http://mingyuanzhou.github.io/STA371G

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 variables and one or more explanatory or predictor variables

Regression in Business

- Optimal portfolio choice:

  - Predict future joint distribution of asset returns

  - Construct 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 future probability of default using known characteristics of borrower

  - Decide whether or not to lend (and if so, how much)

- Auto/health/house insurance:

  - Predict the number and amount of issuance claims

  - Determine insurance premiums

Why?

Straight prediction questions:

- For how much will my house sell?
- How many runs per game will the Red Sox score in 2011?
- How much money will I make by purchasing notes at Lending Club (a peer-to-peer lending platform)?

Explanation and understanding:

- What is the impact of MBA on income?
- How does the returns of a mutual fund relate to the market?
- Does Walmart discriminates against women regarding salaries?
- Does a note of $30,000 issued at Lending Club has a lower probability to default than a note of $3000?

1st Example: Predicting House Prices

Problem:

- Predict market price based on observed characteristics

Solution:

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

Predicting House Prices

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
  - school district, crime rate
  - public transportation
  - traffic noise
- Easy to quantify price and size but what about other variables such as aesthetics, workmanship, etc?

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

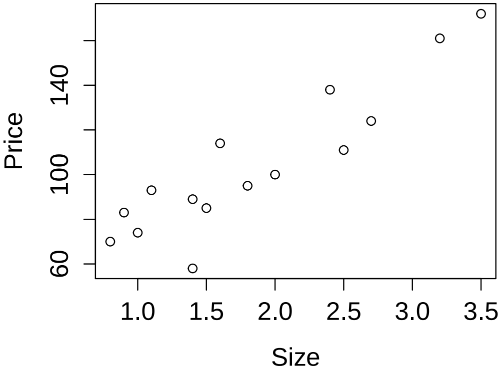
Predicting House Prices

It is much more useful to look at a scatterplot (Using R)

Size=c(0.8,0.9,1.0,1.1,1.4,1.4,1.5,1.6,1.8,2.0,2.4,2.5,2.7,3.2,3.5)

Price=c(70,83,74,93,89,58, 85,114, 95,100,138,111,124,161,172)

plot(Size,Price)



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

Regression Model

$Y =$ response or outcome variable

$X1, X2, X3, \ldots, Xp =$ explanatory or input variables

The general relationship approximated by:

$$Y = f(X_1, X_2, \ldots, X_p) + e$$

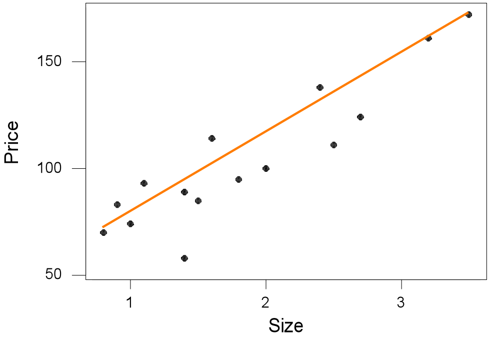
And a linear relationship is written as

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p + e$$

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:

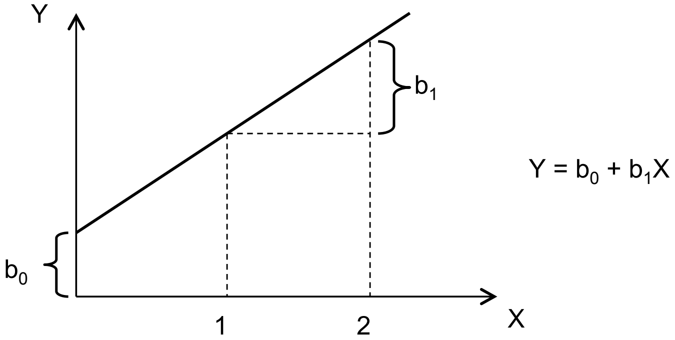
$$Y = b_0 + b_1 X$$

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



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)

Linear Prediction

Can we do better than the eyeball method?

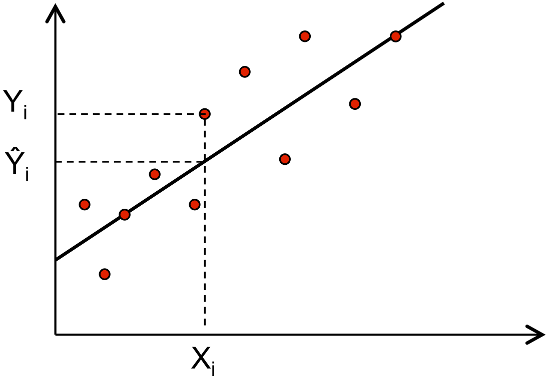
We desire a strategy for estimating the slope and intercept parameters in the model $\hat{Y} = b_0 + b_1 X$

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

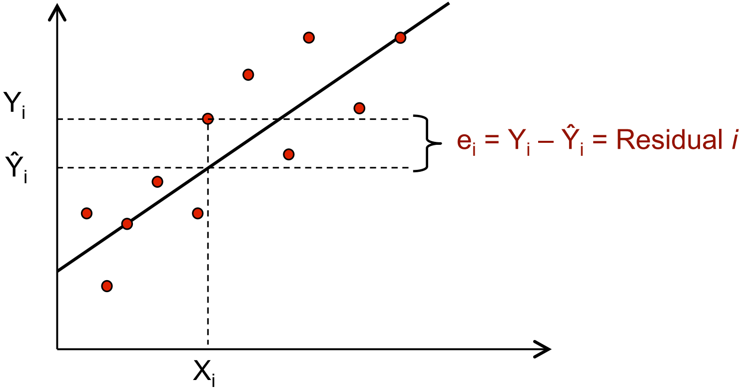
What is the "fitted value"?



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

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.

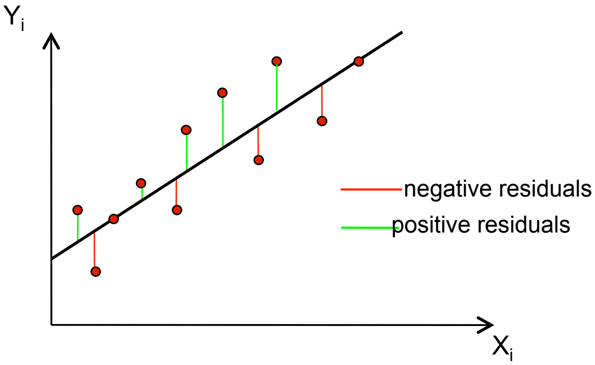
The line fitting process:

- 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



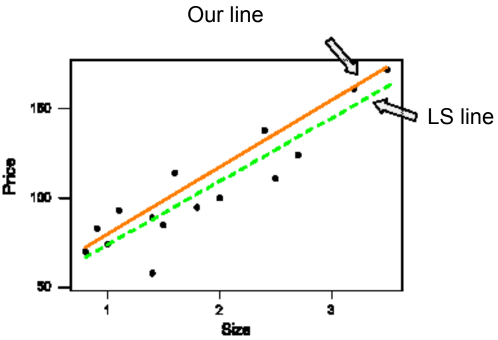
Choose the line to minimize the sum of the squares of the residuals,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - [b_0 + b_1 x_i])^2$$

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?



Eyeball vs. LS Residuals

- eyeball: $b_0 = 35$, $b_1 = 40$
- LS: $b_0 = 38.88$, $b_1 = 35.39$

Size	Price	yhat-eyeball	yhat-LS		e-eyeball	e-LS	e2-eyeball	e2-LS
0.80	70	67	67.19		3.00	2.81	9.00	7.88
0.90	83	71	70.73		12.00	12.27	144.00	150.51
1.00	74	75	74.27		-1.00	-0.27	1.00	0.07
1.10	93	79	77.81		14.00	15.19	196.00	230.76
1.40	89	91	88.42		-2.00	0.58	4.00	0.33
1.40	58	91	88.42		-33.00	-30.42	1089.00	925.67
1.50	85	95	91.96		-10.00	-6.96	100.00	48.49
1.60	114	99	95.50		15.00	18.50	225.00	342.17
1.80	95	107	102.58		-12.00	-7.58	144.00	57.44
2.00	100	115	109.66		-15.00	-9.66	225.00	93.25
2.40	138	131	123.81		7.00	14.19	49.00	201.33
2.50	111	135	127.35		-24.00	-16.35	576.00	267.30
2.70	124	143	134.43		-19.00	-10.43	361.00	108.71
3.20	161	163	152.12		-2.00	8.88	4.00	78.86
3.50	172	175	162.74		-3.00	9.26	9.00	85.84
				sum	-70.00	0.00	3136.00	2598.63

Least Squares – 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

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%*
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

Least Squares – R Output

Size=c(0.8,0.9,1.0,1.1,1.4,1.4,1.5,1.6,1.8,2.0,2.4,2.5,2.7,3.2,3.5)

Price=c(70,83,74,93,89,58, 85,114, 95,100,138,111,124,161,172)

HouseFit=lm(formula=Price∼Size)

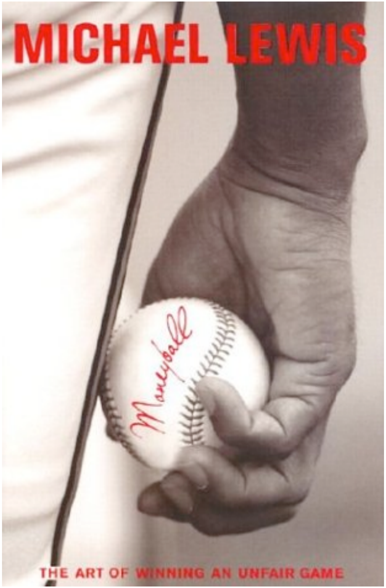
summary(HouseFit)

```
Call:
lm(formula = Price ~ Size)

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
```

2nd Example: Offensive Performance in Baseball

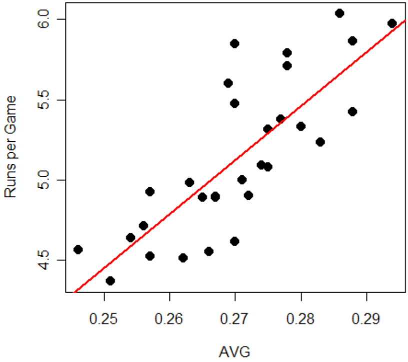
1. Problems:
   - Evaluate/compare traditional measures of offensive performance
   - Help evaluate the worth of a player
2. Solutions:
   - Compare *prediction rules* that forecast runs as a function of either AVG (batting average), SLG (slugging percentage) or OBP (on base percentage)

2nd Example: Offensive Performance in Baseball



Baseball Data – Using AVG

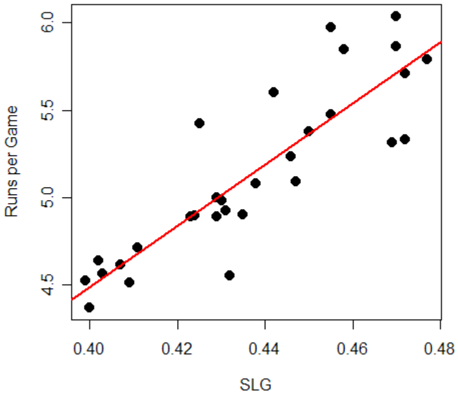
Each observation corresponds to a team in MLB. Each quantity is the average over a season.



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

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

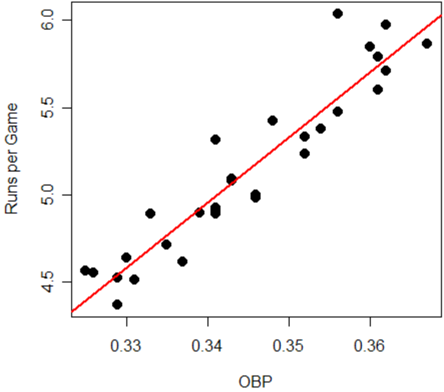
Baseball Data – Using SLG



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

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

Baseball Data – Using OBP



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

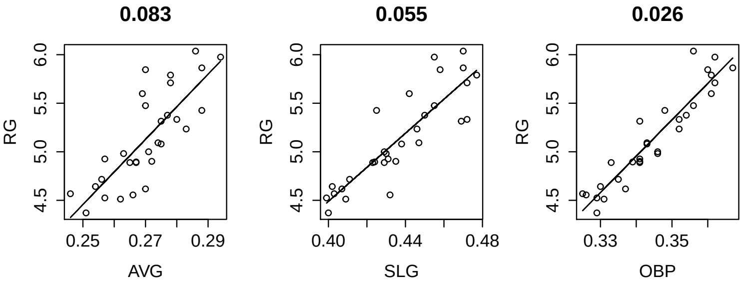
LS fit: Runs/Game = -7.78 + 37.46 OBP

Baseball Data

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

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

Place your Money on OBP!!!

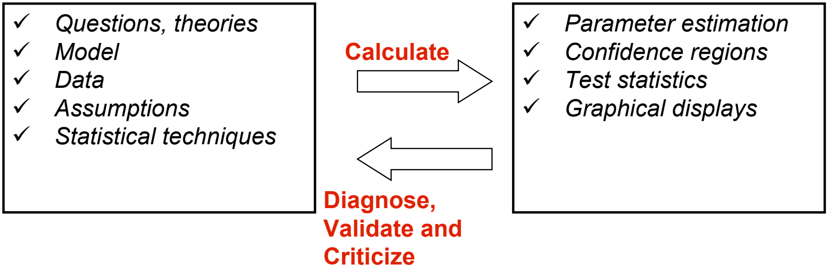


	Average Squared Error
AVG	0.083
SLG	0.055
OBP	0.026

General Steps in Regression Analysis

1. State the problem.
2. Select potentially relevant variables
3. Data collection
4. Model specification
5. Model fitting
6. Model validation and criticism
7. Answering the posed questions

General Regression Strategy

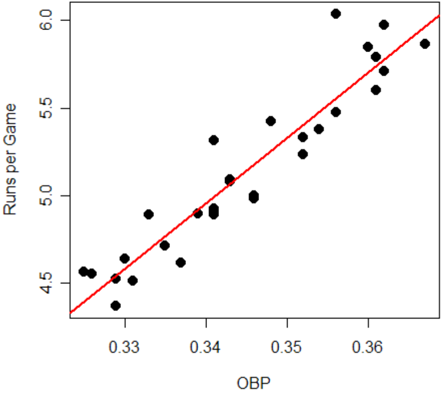


What's next...

The goal in the class is to answer the following questions:

- How do we select a best fitting line?
- Can we say something about the accuracy of our predictions?
- How do we validate and criticize the model?
- How can we use regression analysis to answer relevant business questions?

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 (Optional)

The regression coefficients b_0 and b_1 that minimize

$$SSE(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - [b_0 + b_1 x_i])^2$$

can be calculated as

$$\frac{\partial SSE(b_0, b_1)}{\partial b_0} = 0 \quad \Rightarrow \quad b_0 = \bar{y} - b_1 \bar{x}$$

where $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ and $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

$$\frac{\partial SSE(b_0, b_1)}{\partial b_1} = 0 \quad \Rightarrow \quad b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

The Least Squares Criterion

The regression coefficients b_0 and b_1 that minimize

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - [b_0 + b_1 x_i])^2$$

can be calculated as

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

where,

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