Section 2.1: Intro to Simple Linear Regression & Least Squares

Jared S. Murray
The University of Texas at Austin
McCombs School of Business
Suggested reading: OpenIntro Statistics, Chapter 7.1, 7.2

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 variables

Why?

Straight prediction questions:

- For how much will my house sell?
- How many runs per game will the Red Sox score this year?
- Will this person like that movie? (e.g., Netflix)

Explanation and understanding:

- What is the impact of getting an MBA on lifetime income?
- How do the returns of a mutual fund relate to the market?
- Does Walmart discriminate against women when setting salaries?

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
  - 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, e.g. the price of the house (thousands of dollars)

The variable that we use to aid in prediction is the
independent, explanatory, or input variable, and this is labelled

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

Predicting House Prices

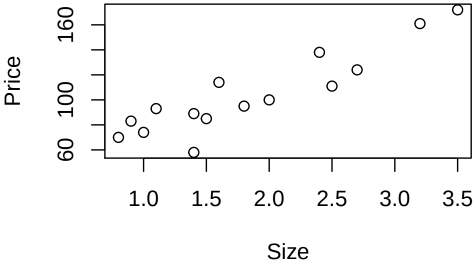
What does this data look like?

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:

```
plot(Price ~ Size, data = housing)
```

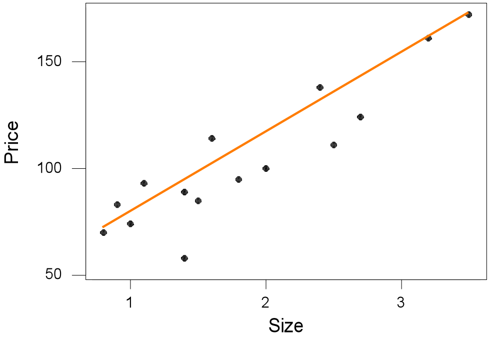


In other words, view the data as points in the $X \times Y$ plane.

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:

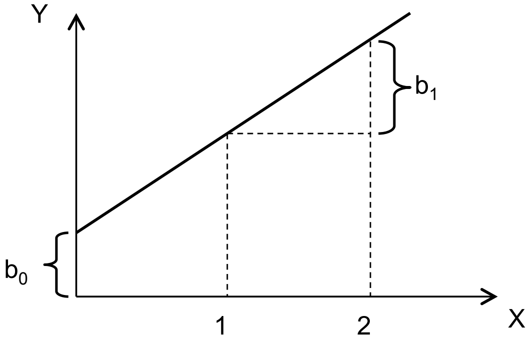
$$Y = b_0 + b_1 X$$

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



Y = b_0 + b_1X

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

Linear Prediction

Can we do better than the eyeball method?

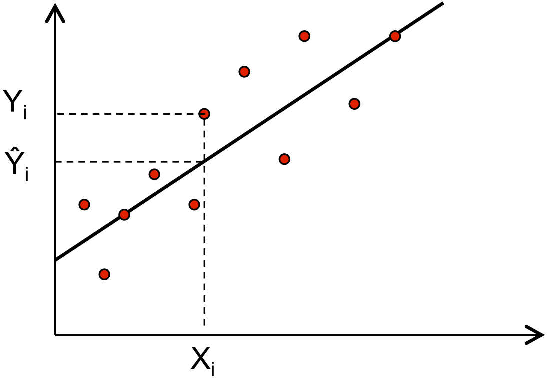
We desire a strategy for estimating the slope and intercept parameters in the model $\hat{Y} = b_0 + b_1 X$

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

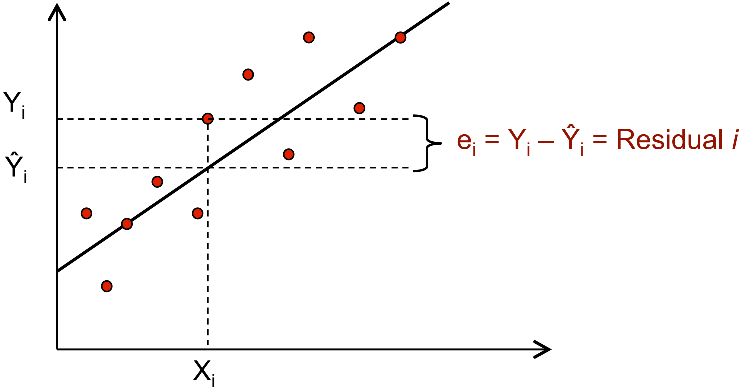
What is the "fitted value"?



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

Linear Prediction

What is the "residual"' for the ith 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:

- Take each residual e_i and assign it a weight e_i^2. Bigger residuals = bigger "mistakes" = higher weights
- Minimize the total of these weights to get best possible fit.

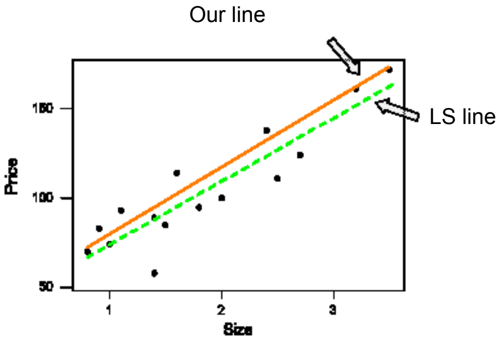
Least Squares chooses b_0 and b_1 to minimize $\sum_{i=1}^{N} e_i^2$

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

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 in R

The `lm` command fits linear (regression) models

```
fit = lm(Price ~ Size, data = housing)
print(fit)

##
## Call:
## lm(formula = Price ~ Size, data = housing)
##
## Coefficients:
## (Intercept)         Size
##       38.88        35.39
```

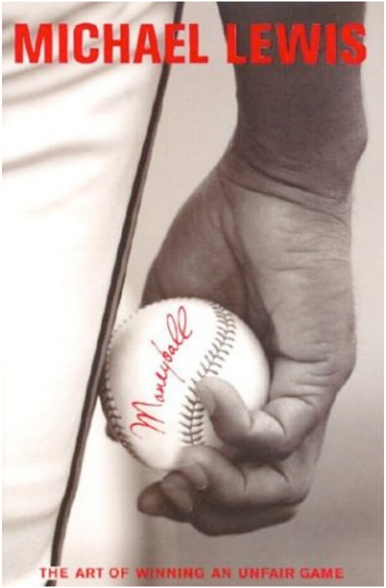
```
fit = lm(Price ~ Size, data = housing)
summary(fit)

##
## Call:
## lm(formula = Price ~ Size, data = housing)
##
## 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

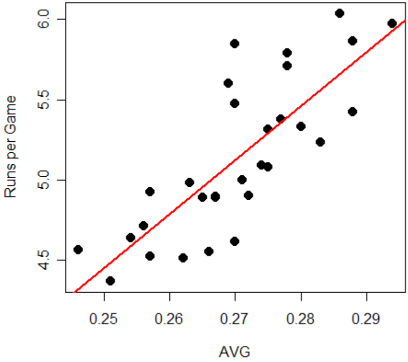
1. Problems:
   - Evaluate/compare traditional measures of offensive performance
   - Help evaluate the worth of a player
2. Solutions:
   - Compare *prediction rules* that forecast runs as a function of either AVG (batting average), SLG (slugging percentage) or OBP (on base percentage)

2nd Example: Offensive Performance in Baseball



Baseball Data – Using AVG

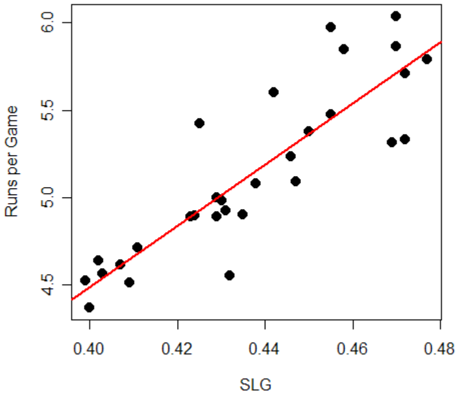
Each observation corresponds to a team in MLB. Each quantity is the average over a season.



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

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

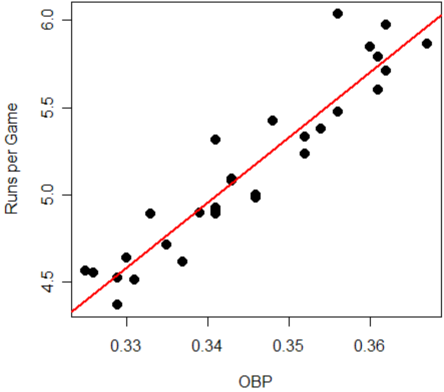
Baseball Data – Using SLG



- 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

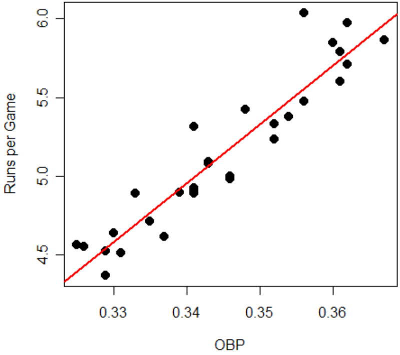
- What is the best prediction rule?
- Let's compare the predictive ability of each model using the average squared error

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - Y_i\right)^2}{n}$$

Place your Money on OBP!

	Average Squared Error
AVG	0.083
SLG	0.055
OBP	0.026

Linear Prediction



$$\hat{Y}_{n+1} = b_0 + b_1 x_{n+1}$$

- b_0 is the intercept and b_1 is the slope
- We find b_0 and b_1 using *Least Squares*
- For a new value of the independent variable OBP (say x_{n+1}) we can predict the response Y_{n+1} using the fitted line

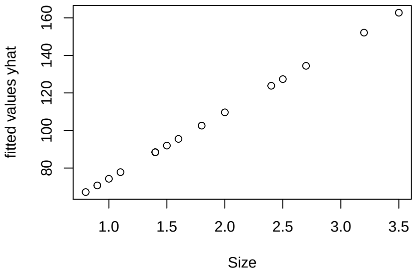
More on Least Squares

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...

The Fitted Values and X

```
plot(predict(fit) ~ Size, data = housing, ylab = "fitted va
```

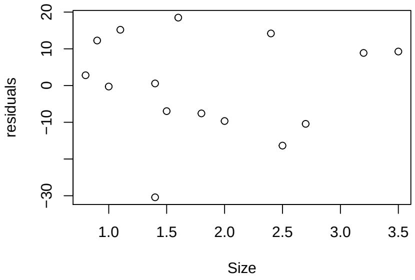



```
cor(predict(fit), housing$Size)

## [1] 1
```

The Residuals and X

```r
plot(resid(fit) ~ Size, data = housing, ylab = "residuals")
```




```r
mean(resid(fit));  cor(resid(fit), housing$Size)

## [1] -9.633498e-17
## [1] 2.120636e-17
```

(i.e., zero). What's going on here? Next time....