Regression

Communication Research Methods

Jennifer Pan

Assistant Professor
Department of Communication
Stanford University

February 19, 2016

Announcements

- Monday Feb 29 Section and Tuesday Mar 1 Office Hours
- Data for class next Wed

Where we are

- Previously
  - Describing data
  - Assessing hypotheses by making comparisons
  - Making comparisons with interval variables (correlation)
- Today
  - Thinking systematically about how X relates to Y (linear regression)

16 NBA Players in 2007-008

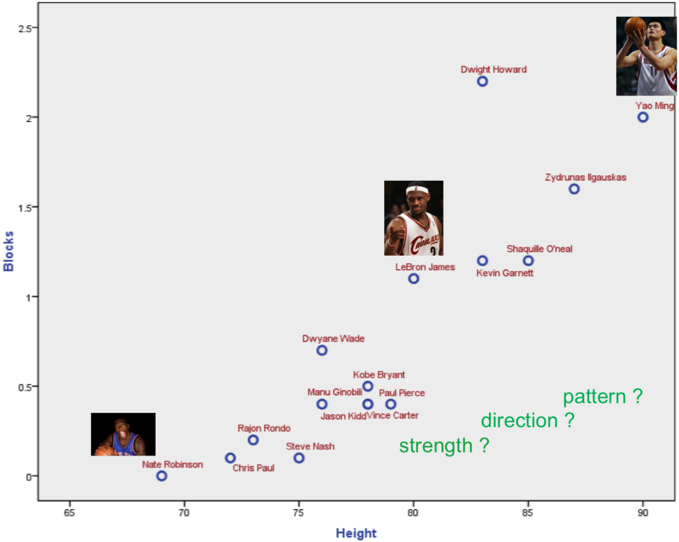
Name	height	weight	age	rebound	blocks
Nate Robinson	69	180	23	3.1	0
Chris Paul	72	175	22	4	0.1
Rajon Rondo	73	171	21	4.2	0.2
Steve Nash	75	178	33	3.5	0.1
Dwyane Wade	76	216	25	4.2	0.7
Jason Kidd	76	210	34	6.5	0.4
Vince Carter	78	220	30	6	0.4
Kobe Bryant	78	205	29	6.3	0.5
Manu Ginobili	78	205	30	4.8	0.4
Paul Pierce	79	235	30	5.1	0.4
LeBron James	80	250	23	7.9	1.1
Dwight Howard	83	265	22	14.2	2.2
Kevin Garnett	83	253	31	9.2	1.2
Shaquille O'neal	85	325	35	10.6	1.2
Zydrunas Ilgauskas	87	260	32	9.3	1.6
Yao Ming	90	310	27	10.8	2

We Might Ask...

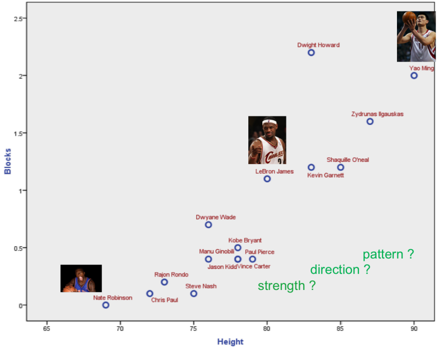
how is height (Y) related to average number of blocks per game (X) for NBA players?

Name	height	weight	age	rebound	blocks
Nate Robinson	[illegible]	180	23	3.1	0
Chris Paul	72	[illegible]	22	4	0.1
Rajon Rondo	73	[illegible]	[illegible]	4.2	0.2
Steve Nash	75	178	[illegible]	3.5	0.1
Dwyane Wade	76	216	25	[illegible]	0.7
Jason Kidd	76	210	34	6.5	[illegible]
Vince Carter	78	220	30	6	0.4
Kobe Bryant	78	205	29	6.3	0.5
Manu Ginobili	78	205	30	4.8	0.4
Paul Pierce	79	235	30	5.1	0.4
LeBron James	80	250	23	7.9	1.1
Dwight Howard	83	265	22	14.2	[illegible]
Kevin Garnett	83	253	31	9.2	[illegible]
Shaquille O'neal	85	325	35	10.6	1.2
Zydrunas Ilgauskas	87	260	32	[illegible]	1.6
Yao Ming	90	310	27	[illegible]	2

We Can Make a Scatterplot



We Can Calculate Correlation



- `cor(height, blocks)` = 0.88
- Strong, positive (linear) association
- Being tall is associated with blocking more
- Being short is associated with blocking less

We Want to Know More!

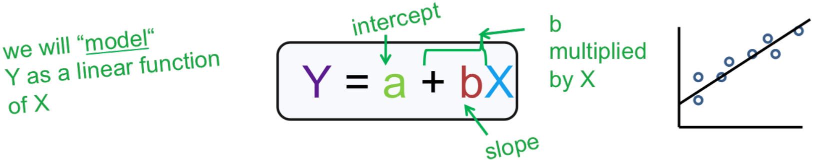
We might want to ask:

1. On average, what is the effect of a one inch increase in a player's height on his blocks per game?
2. If we have a NBA player and know his height, what is his predicted blocks per game?

Correlation cannot tell us Regression can tell us

Linear Regression

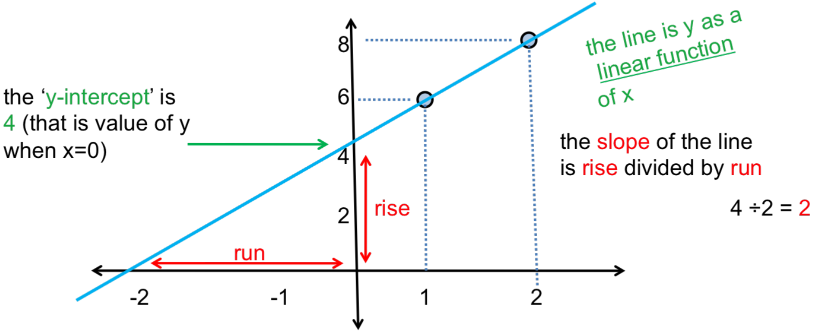
- We will make our Y depend on our X in the following way:



- This is called a "linear" function because it produces a straight line graph between X and Y
- For any value of X (and a) we will get an expected value for Y
- For any value of height (and a) we will get an expected value for blocks
- Finding the values for b and a amounts to drawing a 'best fit' line for the scatter plot

Linear Regression: Back to High School

- Suppose $x_1 = 1$, $y_1 = 6$ and $x_2 = 2$, $y_2 = 8$
- What is the relationship between them?



- y-intercept $= 4$, slope $= 2$

Linear Regression: Back to High School

- Suppose $x_1 = 1$, $y_1 = 6$ and $x_2 = 2$, $y_2 = 8$
- y-intercept $= 4$, slope $= 2$

y value= intercept + (slope times X value)

$$y = 4 + 2X$$

$$6 = 4 + (2 \times 1)$$

$$8 = 4 + (2 \times 2)$$

$$Y = a + bX$$

Linear Regression: Notation

When we fit the equation to our data...

$$Y = a + bX$$

...we will be *estimating* the a and b in the population.

That population relationship is written with Greek letters:

$$\boxed{Y = \alpha + \beta X}$$

Our *estimates* from the sample are often written with 'hats':

"alpha hat" $\hat{\alpha}$ $\hat{\beta}$ "beta hat"

We use these terms:

constant coefficient

Back to our NBA Example

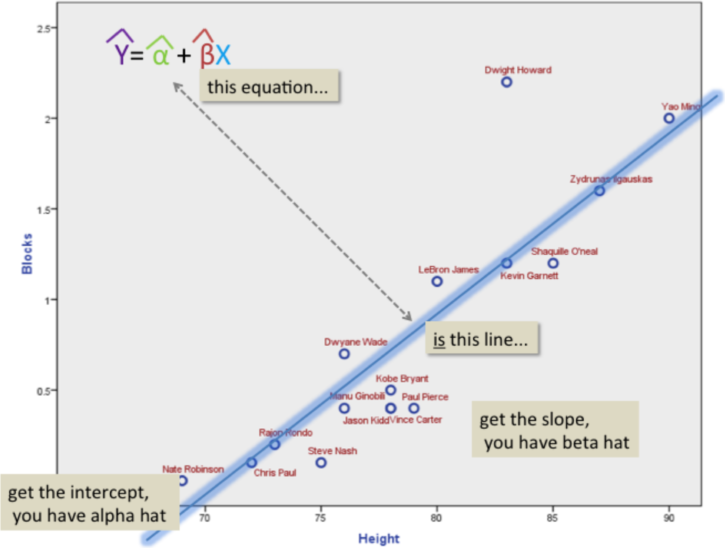
We want to know:

1. On average, what is the effect of a one inch increase in a player's height on his blocks per game?
2. If we have a NBA player and know his height, what is his predicted blocks per game?

Regression can tell us:

- the 'best fit' (straight) line for the data
- the equation for that line
- a predicted Y for any value of X (a predicted blocks per game for any height)

Back to our NBA Example



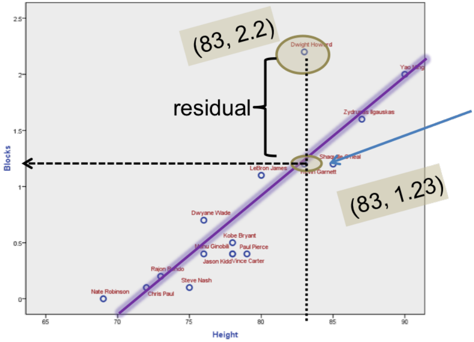
Back to our NBA Example

- $\hat{\alpha} = -7.734$ (the constant)...
  - is the intercept on the y-axis
  - tells us what value Y takes when X (height) is zero
- $\hat{\beta} = 0.108$ (the coefficient)...
  - average change in Y (blocks) for a one unit increase in X (height)
  - coefficient are always in terms of Y's units!
- Interpretation of $\hat{\beta}$
  - On average...Y changes by [$\hat{\beta}$][Y's units] for a 1 [X's unit] increase in X
  - On average...a player's blocks per game changes by [0.108][blocks] for a 1 [inch] increase in height

Making Predictions

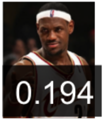
- To predict Y for any X, just plug the X into the equation and calculate $\hat{y}$
- $\hat{y} = \hat{\alpha} + \hat{\beta}x$
- $\hat{y}$ is the predicted value
- NBA example:
  - $\hat{y} = -7.734 + 0.108x$
  - Tony Parker is 6'2" (74in) (not in sample)
  - $\hat{y} = -7.734 + 0.108(74) = 0.258$
  - Parker's actual average was 0.1 blocks

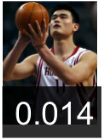
Actual (y) vs. Predicted ($\hat{y}$)

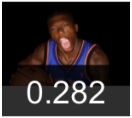


- y is the actual value of y in the sample
- $\hat{y}$ is what we predict based on the regression line
- Example: Dwight Howard
  - Howard has 2.2 actual blocks per game
  - Howard is predicted to have 1.23 from our regression
  - Difference between actual value and predicted value is the residual
  - Residual for Howard is 2.2 – 1.23 = 0.97

Calculating the Line







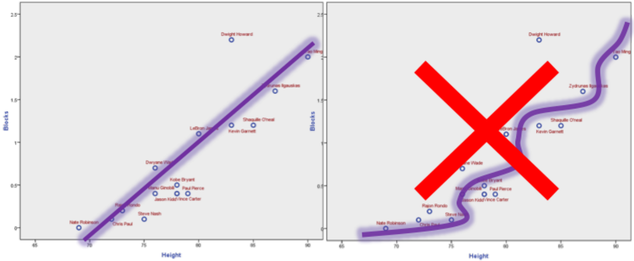
- Residual = actual - predicted = $y - \hat{y}$
- We have a residual for each observation
- We draw the line so it minimizes the residuals
- Specifically, minimize the sum of squared residuals
- OLS (ordinary least squares) estimation

Linear Regression

- For a given value of X (e.g., height of 67in, 69in)...Y (blocks) can take different values
- We will focus on the mean value Y takes for a given value of X

$$\text{average}(Y|X) = \alpha + \beta X$$

- The slope β is always the same...this is a linear regression model

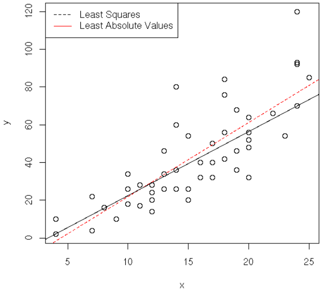


Ordinary Least Squares

- In theory, we could use any equation to link y to x:

$$\text{average}(Y|X) = f(X)$$

- Ordinary Least Squares (estimation) is the technique we use to draw the line through the points
- OLS is so common it's often synonymous with linear regression
- But, there are other ways to draw a straight line through points (least absolute values)

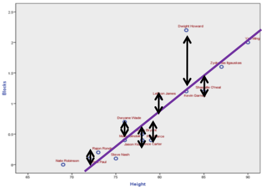


Ordinary Least Squares

the line we draw...

"least" "squares"

will *minimize* the total squared vertical distances from the *actual* data points in the sample



minimize the sum of

the *squared residuals*

Linear Regression

- Units matter (correlation, unit-less)
  - The effect of average temperature (X) on ice cream sales (Y)
  - $\hat{\beta}$ will change if we measure temp in Celcius vs in Fahrenheit (and sales in dollars vs in euros)
  - Remember: $\hat{\alpha}$ and $\hat{\beta}$ are in y's units...not in x's units
- Asymmetric (correlation, symmetric)
  - Regression of Y on X is...not the same as regression of X on Y

Ordinary Least Squares Linear Regression in R

`lm()`