Unit – III
DESCRIBING RELATIONSHIPS

Correlation – Scatter plots – correlation coefficient for quantitative data – computational formula for correlationcoefficient – Regression – regression line – least squares regression line – Standard error of estimate –interpretation of r2 – multiple regression equations – regression towards the mean

Correlation

Correlation refers to a process for establishing the relationships between two variables. You learned a way to get a general idea about whether or not two variables are related, is to plot them on a "scatter plot". While there are many measures of association for variables which are measured at the ordinal or higher level of measurement, correlation is the most commonly used approach.

Types of Correlation

- **Positive Correlation** – when the values of the two variables move in the same direction so that an increase/decrease in the value of one variable is followed by an increase/decrease in the value of the other variable.
- **Negative Correlation** – when the values of the two variables move in the opposite direction so that an increase/decrease in the value of one variable is followed by decrease/increase in the value of the other variable.
- **No Correlation** – when there is no linear dependence or no relation between the two variables.

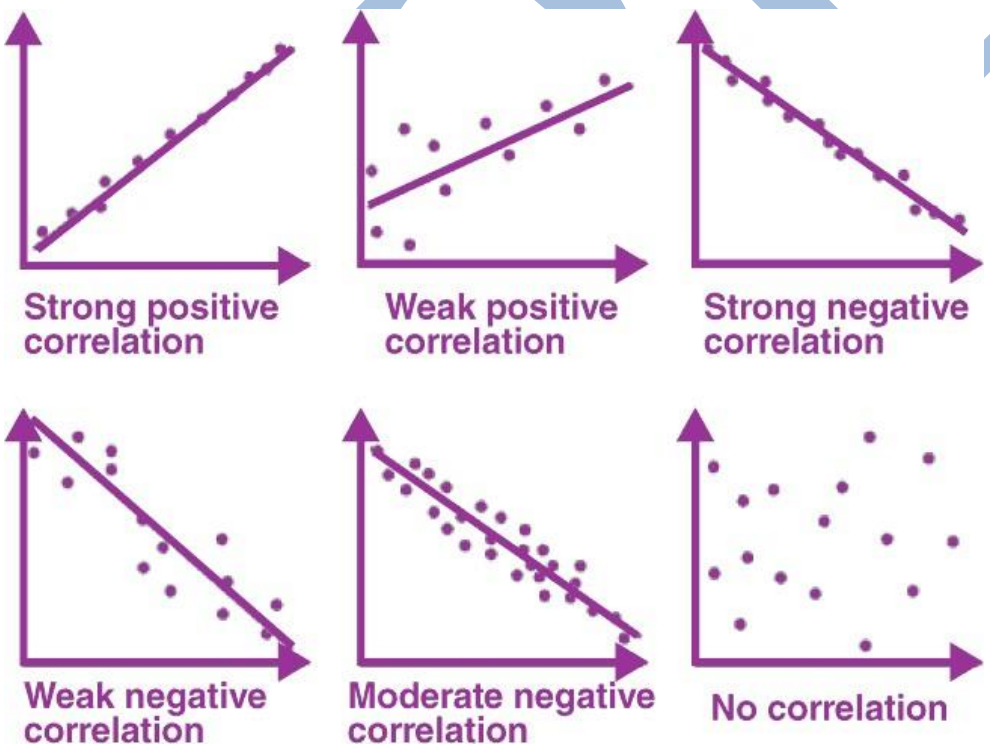


Table 6.2
THREE TYPES OF RELATIONSHIPS

A. POSITIVE RELATIONSHIP

FRIEND	SENT	RECEIVED
Doris	13	14
Steve	9	18
Mike	7	12
Andrea	5	10
John	1	6

B. NEGATIVE RELATIONSHIP

FRIEND	SENT	RECEIVED
Doris	13	6
Steve	9	10
Mike	7	14
Andrea	5	12
John	1	18

C. LITTLE OR NO RELATIONSHIP

FRIEND	SENT	RECEIVED
Doris	13	10
Steve	9	18
Mike	7	12
Andrea	5	6
John	1	14

SCATTERPLOTS

A scatter plot is a graph containing a cluster of dots that represents all pairs of scores. In other words Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents datapoints on a two-dimensional plane or on a Cartesian system.

Construction of scatter plots

- The independent variable or attribute is plotted on the X-axis.
- The dependent variable is plotted on the Y-axis.

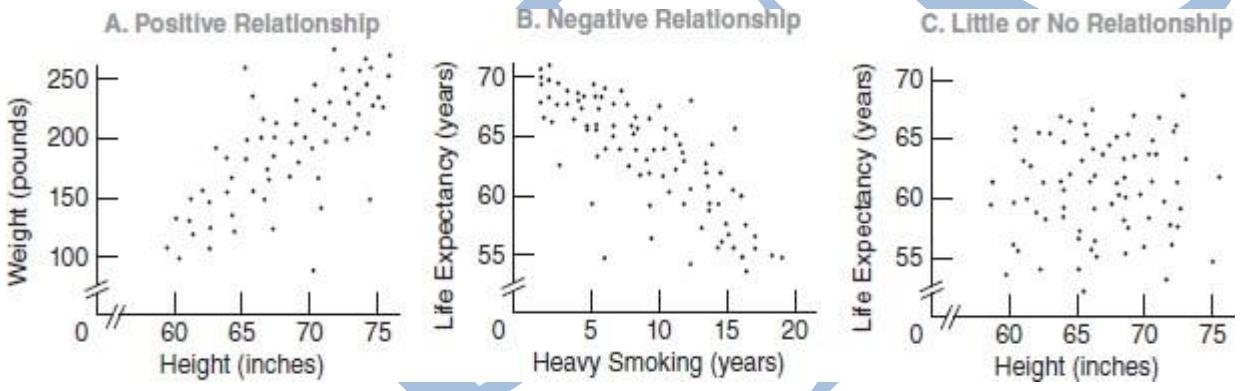
- Use each pair of scores to locate a dot within the scatter plot

Positive, Negative, or Little or No Relationship?

The first step is to note the tilt or slope, if any, of a dot cluster.
A dot cluster that has a slope from the lower left to the upper right, as in panel A of below figure reflects a **positive relationship.**

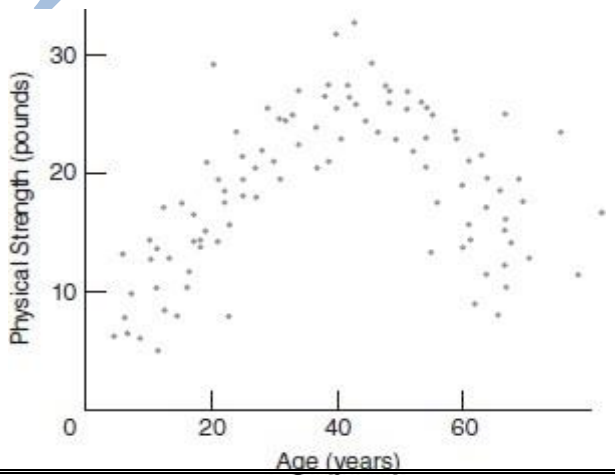
A dot cluster that has a slope from the upper left to the lower right, as in panel B of below figure reflects a **negative relationship.**

A dot cluster that lacks any apparent slope, as in panel C of below figure reflects **little or no relationship.**



Perfect Relationship
A dot cluster that equals (rather than merely approximates) a straight line reflects a perfect relationship between two variables.

Curvilinear Relationship
The previous discussion assumes that a dot cluster approximates a straight line and, therefore, reflects a **linear relationship.** But this is not always the case. Sometimes a dot cluster approximates a bent or curved line, as in below figure, and therefore reflects a **curvilinear relationship.**



A CORRELATION COEFFICIENT FOR QUANTITATIVE DATA : *r*

The correlation coefficient, r, is a summary measure that describes the extent of the statistical relationship between two interval or ratio level variables.

Properties of r

- The correlation coefficient is scaled so that it is always between -1 and +1.
- When r is close to 0 this means that there is little relationship between the variables and the farther away from 0 r is, in either the positive or negative direction, the greater the relationship between the two variables.
- The sign of r indicates the type of linear relationship, whether positive or negative.
- The numerical value of r, without regard to sign, indicates the strength of the linear relationship.
- A number with a plus sign (or no sign) indicates a positive relationship, and a number with a minus sign indicates a negative relationship

COMPUTATION FORMULA FOR *r*

Calculate a value for r by using the following computation formula:

CORRELATION COEFFICIENT (COMPUTATION FORMULA)

$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$$

Where the two sum of squares terms in the denominator are defined as

$$SS_x = \Sigma\left(X - \bar{X}\right)^2 = \Sigma X^2 - \frac{\left(\Sigma X\right)^2}{n}$$

$$SS_y = \Sigma\left(Y - \bar{Y}\right)^2 = \Sigma Y^2 - \frac{\left(\Sigma Y\right)^2}{n}$$

The sum of the products term in the numerator, *SPxy*, is defined in below formula

$$SP_{xy} = \Sigma\left(X - \bar{X}\right)\left(Y - \bar{Y}\right) = \Sigma XY - \frac{\left(\Sigma X\right)\left(\Sigma Y\right)}{n}$$

Or the formula is written as

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where n = Number of Information

Σx = Total of the First Variable Value

Σy = Total of the Second Variable

Value

Σxy = Sum of the Product of first & Second Value

Σx^2 = Sum of the Squares of the First Value

Σy^2 = Sum of the Squares of the Second Value

Table 6.3
CALCULATION OF *r*: COMPUTATION FORMULA

A. COMPUTATIONAL SEQUENCE

Assign a value to n (1), representing the number of pairs of scores.
Sum all scores for X (2) and for Y (3).
Find the product of each pair of X and Y scores (4), one at a time, then add all of these products (5).
Square each X score (6), one at a time, then add all squared X scores (7).
Square each Y score (8), one at a time, then add all squared Y scores (9).
Substitute numbers into formulas (10) and solve for SP_{xy}, SS_x, and SS_y.
Substitute into formula (11) and solve for r.

B. DATA AND COMPUTATIONS

	CARDS		4	6	8
FRIEND	**SENT, *X***	**RECEIVED, *Y***	***XY***	***X*²**	***Y*²**
Doris	13	14	182	169	196
Steve	9	18	162	81	324
Mike	7	12	84	49	144
Andrea	5	10	50	25	100
John	1	6	6	1	36

1 $n = 5$ 2 $\Sigma X = 35$ 3 $\Sigma Y = 60$ 5 $\Sigma XY = 484$ 7 $\Sigma X^2 = 325$ 9 $\Sigma Y^2 = 800$

10 $$SP_{xy} = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n} = 484 - \frac{(35)(60)}{5} = 484 - 420 = 64$$

$$SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 325 - \frac{(35)^2}{5} = 325 - 245 = 80$$

$$SS_y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} = 800 - \frac{(60)^2}{5} = 800 - 720 = 80$$

11 $$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} = \frac{64}{\sqrt{(80)(80)}} = \frac{64}{80} = .80$$

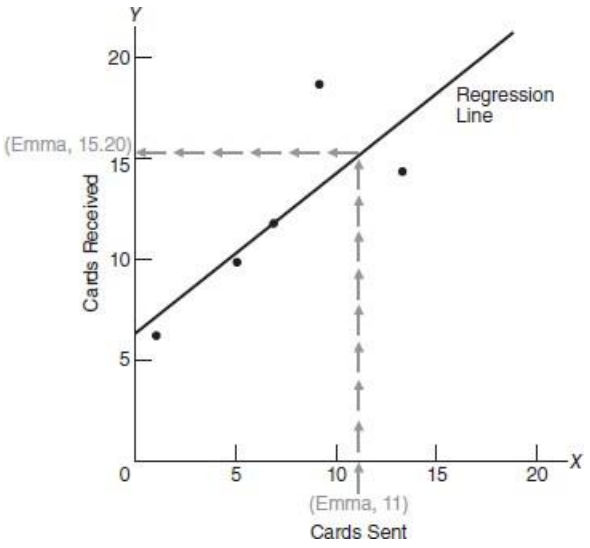
REGRESSION

A regression is a statistical technique that relates a dependent variable to one or more independent (explanatory) variables. A regression model is able to show whether changes observed in the dependent variableare associated with changes in one or more of the explanatory variables.

Regression captures the correlation between variables observed in a data set, and quantifies whether those correlations are statistically significant or not.

A Regression Line

a regression line is a line that best describes the behaviour of a set of data. In other words, it's a line that bestfits the trend of a given data.



The purpose of the line is to describe the interrelation of a dependent variable (Y variable) with one or many independent variables (X variable). By using the equation obtained from the regression line an analyst can forecast future behaviours of the dependent variable by inputting different values for the independent ones.

Types of regression

The two basic types of regression are

- Simple linear regression

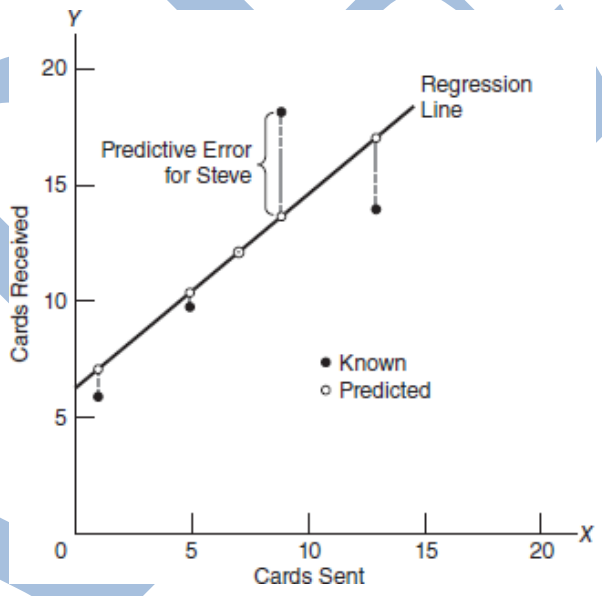
Simple linear regression uses one independent variable toexplain or predict the outcome of the dependent variable Y

- Multiple linear regression

Multiple linear regressions use two or more independentvariables to predict the outcome

Predictive Errors

Prediction error refers to the difference between the predicted values made by some model and theactual values.



LEAST SQUARES REGRESSION LINE

The placement of the regression line minimizes not the total predictive error but the total squared predictive error, that is, the total for all squared predictive errors. When located in this fashion, the regression line is often referred to as the least squares regression line.

The Least Squares Regression Line is the line that minimizes the sum of the residuals squared. The residual is the vertical distance between the observed point and the predicted point, and it is calculated by subtracting ŷ from y.

Formula

$y' = bx+a$ b – slope , a – y intercept $b = \frac{N\Sigma(xy) - \Sigma x \Sigma y}{N\Sigma(x^2) - (\Sigma x)^2}$

$b = \frac{\Sigma y - m \Sigma x}{N}$

Example

"x"	"y"
2	4
3	5
5	7
7	10
9	15

Step 1: For each (x,y) calculate x^2 and xy:

x	**y**	**x^2**	**xy**
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135

Step 2: Sum x, y, x2 and xy (gives us Σx, Σy, Σx2 and Σxy):

Σx: 26 Σy: 41 Σx^2: 168 Σxy: 263

Step 3: Calculate Slope b

$$b = \frac{N \Sigma(xy) - \Sigma x \Sigma y}{N \Sigma(x^2) - (\Sigma x)^2}$$

$$= \frac{5 \times 263 - 26 \times 41}{5 \times 168 - 262}$$

$$= \frac{1315 - 1066}{840 - 676}$$

$$= \frac{249}{164}$$

$b = 1.5183.$

Step 4: Calculate Intercept

$$aa = \frac{\Sigma y - b \Sigma x}{N}$$

$$= \frac{41 - 1.5183 \times 26}{5}$$

$a = 0.3049.$

Step 5: $y' = bx+a$

$y' = 1.518x + 0.305$

APEC

x	y	y = 1.518x + 0.305	error
2	4	3.34	−0.66
3	5	4.86	−0.14
5	7	7.89	0.89
7	10	10.93	0.93
9	15	13.97	−1.03

To predict the y value we can assume any value for x.Assume x = 8.
Then y = 1.518 x 8 + 0.305
= 12.45

STANDARD ERROR OF ESTIMATE ,$s_{y|x}$

The standard error of the estimate is a measure of the accuracy of predictions. The regression line is the line that minimizes the sum of squared deviations of prediction (also called the sum of squares error), and the standard error of the estimate is the square root of the average squared deviation.

The standard error of estimate and symbolized as $s_{y|x}$, this estimate of predictive error complies with the general format for any sample standard deviation, that is, the square root of a sum of squares term dividedby its degrees of freedom.

$$SS_{y|x} = \sum(Y - Y')^2$$

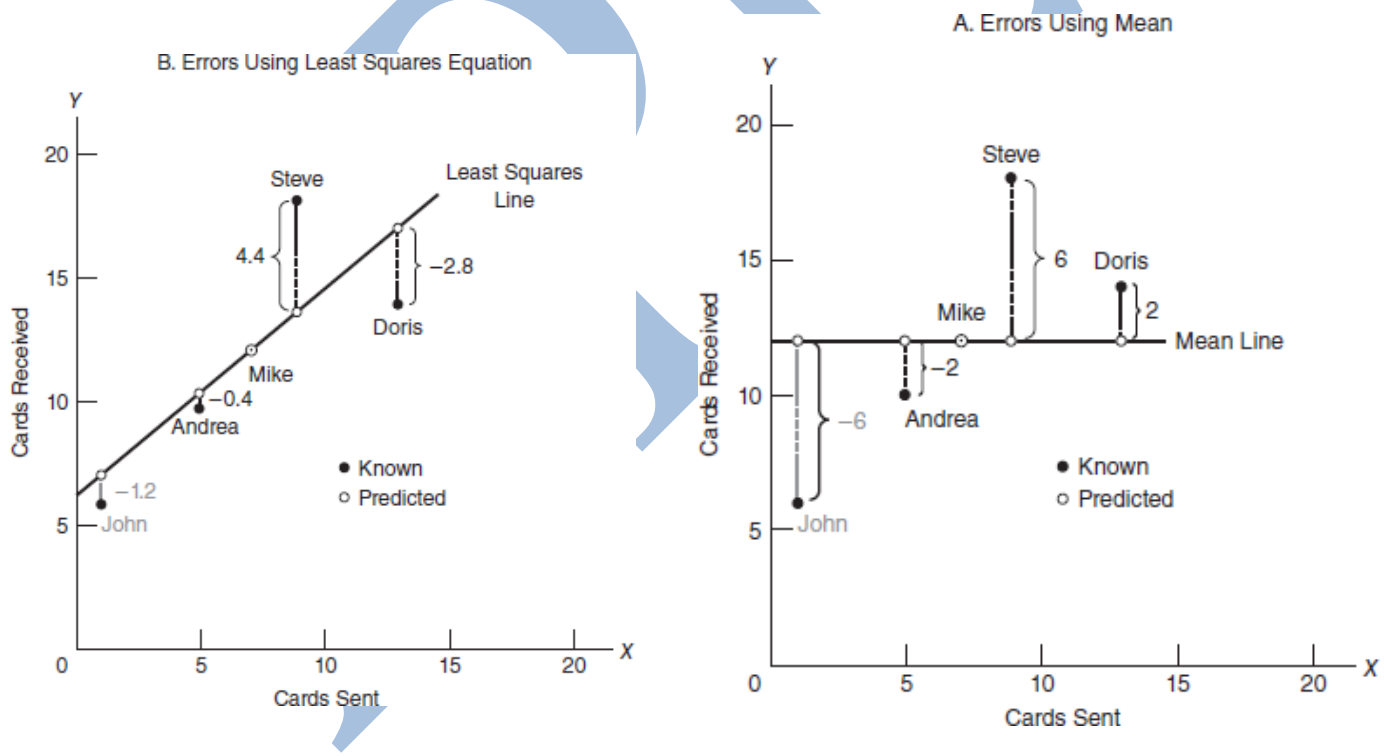


Fig. Predictive errors for five friends

Example

Calculate the standard error of estimate for the given X and Y values. X = 1,2,3,4,5 Y=2,4,5,4,5

Solution
Create five columns labeled x, y, y', y – y', (y – y')2 and N=5

x	y	x^2	xy	Y'= bx+a	y-y'	$(y – y')^2$
1	2	1	2	2.8	-0.8	0.64
2	4	4	8	3.4	0.6	0.36
3	5	9	15	4.0	1	1
4	4	16	16	4.6	-0.6	0.36
5	5	25	25	5.2	-0.2	0.04
Σx:15	**Σy:20**	**$Σx^2$:55**	**Σxy:66**			**$Σ(y – y')^2$ = 2.4**

Note: for finding b value we have to find xy and x^2, so add xy and x^2 column in table

$$b = \frac{N \Sigma(xy) - \Sigma x \Sigma y}{N \Sigma(x^2) - (\Sigma x)^2}$$

$$b = \frac{5(66) - 15x20}{5(55) - (15)^2}$$

$$= \frac{330 - 300}{275 - 225}$$

b= 30/50 = 0.6

$$a = \frac{\Sigma y - b \Sigma x}{N}$$

$$= \frac{20 - (0.6 \times 15)}{5}$$

$$= \frac{20 - 11}{5}$$

a= 9/5 = 2.2

$SSy/x = \sqrt{((y-y')^2 / n-2)}$

$=\sqrt{(2.4/3)}$

SSy/x = 0.894

APEC

INTERPRETATION OF *r* 2

R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

R-squared can take any values between 0 to 1. Although the statistical measure provides some useful insights regarding the regression model, the user should not rely only on the measure in the assessment of a statistical model.

In addition, it does not indicate the correctness of the regression model. Therefore, the user should always draw conclusions about the model by analyzing r-squared together with the other variables in astatistical model.

The most common interpretation of r-squared is how well the regression model explains observed data.

$$r^2 = \frac{SS_{Y'}}{SS_Y} = \frac{SS_Y - SS_{Y|X}}{SS_Y}$$

$$SS_{y'} = \sum(Y' - \bar{Y})^2$$

MULTIPLE REGRESSION EQUATIONS

Multiple regression is a statistical technique applied on datasets dedicated to draw out a relationship between one response or dependent variable and multiple independent variables.

Multiple regression works by considering the values of the available multiple independent variables and predicting the value of one dependent variable.

Example:

A researcher decides to study students' performance from a school over a period of time. He observed that as the lectures proceed to operate online, the performance of students started to decline as well. The parameters forthe dependent variable "decrease in performance" are various independent variables like "lack of attention, more internet addiction, neglecting studies" and much more.

Formula to find multiple regression

$$y = b1x1 + b2x2 + \dots bnxn + a$$

REGRESSION TOWARD THE MEAN

Regression toward the mean refers to a tendency for scores, particularly extreme scores, to shrink toward themean.

In statistics, regression toward the mean (also called reversion to the mean, and reversion to mediocrity) is a concept that refers to the fact that if one sample of a random variable is extreme, the next sampling of the samerandom variable is likely to be closer to its mean.

Example

A military commander has two units return, one with 20% casualties and another with 50% casualties. He praises the first and berates the second. The next time, the two units return with the opposite results. From this experience, he "learns" that praise weakens performance and berating increases performance.

The Regression Fallacy

The regression fallacy is committed whenever regression toward the mean is interpreted as a real, ratherthan a chance, effect.

The regression fallacy can be avoided by splitting the subset of extreme observations into two groups

Table 7.4
REGRESSION TOWARD THE MEAN: BATTING AVERAGES OF TOP 10 HITTERS IN MAJOR LEAGUE BASEBALL DURING 2014 AND HOW THEY FARED DURING 2015

	BATTING AVERAGES*		
TOP 10 HITTERS (2014)	2014	2015	REGRESS TOWARD MEAN?
1. J. Alture	.341	.313	Yes
2. V. Martinez	.335	.282	Yes
3. M. Brantley	.327	.310	Yes
4. A. Beltre	.324	.287	Yes
5. J. Abreu	.317	.290	Yes
6. R. Cano	.314	.287	Yes
7. A. McCutchen	.314	.292	Yes
8. M. Cabrera	.313	.338	No
9. B. Posey	.311	.318	No
10. B. Revere	.306	.306	No