Jacob Strong

Professor Caso

Advanced Placement Statistics

November 19, 2021

Display of Data

*This data is without outliers. It excludes data from David Fletcher and Elvis Andrus.

Player Name	Weight	At-Bats	Homeruns	AB/Hr	Hr/AB
Kelvin Guteirrez	215	272	3	90.66667	0.011029
Mancini	215	556	21	26.47619	0.03777
Mountcastle	185	534	33	16.18182	0.061798
Dalbec	227	417	25	16.68	0.059952
Bogaerts	218	529	23	23	0.043478
J.D Martinez	230	570	28	20.35714	0.049123
Judge	282	550	39	14.10256	0.070909
G. Sanchez	230	383	23	16.65217	0.060052
LeMahieu	220	597	10	59.7	0.01675
Arozarena	185	529	20	26.45	0.037807
Meadows	225	518	27	19.18519	0.052124
Zunino	235	333	33	10.09091	0.099099
T. Hernadez	205	550	32	17.1875	0.058182

Vladimir	250	604	48	12.58333	0.07947
Espinal	175	222	2	111	0.009009
Anderson	190	527	17	31	0.032258
Abbreau	245	566	30	18.86667	0.053004
Vaughn	215	417	15	27.8	0.035971
Reyes	265	418	30	13.93333	0.07177
Zimmer	220	299	8	37.375	0.026756
Rosario	189	550	11	50	0.02
Candelario	221	557	16	34.8125	0.028725
Schoop	225	623	22	28.31818	0.035313
Castro	170	413	9	45.88889	0.021792
Santana	210	565	19	29.73684	0.033628
Taylor	212	483	12	40.25	0.024845
Merifeild	195	664	10	66.4	0.01506
Buxton	190	235	19	12.36842	0.080851
Sano	272	470	30	15.66667	0.06383
Kepler	220	426	19	22.42105	0.044601
Brantley	209	469	8	58.625	0.017058
Tucker	199	506	30	16.86667	0.059289
Diaz	195	294	8	36.75	0.027211
Gosselin	200	345	7	49.28571	0.02029
Ohtani	210	537	46	11.67391	0.085661
Pinter	210	214	6	35.66667	0.028037

Kemp	160	330	8	41.25	0.024242
France	217	571	18	31.72222	0.031524
Fraley	195	214	9	23.77778	0.042056
Murphey	218	277	11	25.18182	0.039711
Lowe	245	557	18	30.94444	0.032316
heim	220	265	10	26.5	0.037736
Garcia	180	581	31	18.74194	0.053356
Riley	240	590	33	17.87879	0.055932
Duvall	215	513	38	13.5	0.074074
Swanson	190	588	27	21.77778	0.045918
Rojas	188	495	9	55	0.018182
Brinson	212	274	9	30.44444	0.032847
Sanchez	190	588	27	21.77778	0.045918
McNeil	195	386	7	55.14286	0.018135
Alonso	245	561	37	15.16216	0.065954
Nimmo	206	325	8	40.625	0.024615
Hoskins	225	389	27	14.40741	0.069409
Harper	220	488	34	14.35294	0.069672
Realmuto	210	476	17	28	0.035714
Gracia	190	236	6	39.33333	0.025424
Soto	224	502	29	17.31034	0.057769
Keibom	190	217	6	36.16667	0.02765
Happ	205	465	25	18.6	0.053763

Contreras	210	413	21	19.66667	0.050847
Ortega	160	296	11	26.90909	0.037162
Farmer	205	483	16	30.1875	0.033126
India	200	532	21	25.33333	0.039474
Suarez	213	505	31	16.29032	0.061386
Vogelbach	250	215	9	23.88889	0.04186
Taylor	185	243	12	20.25	0.049383
Urias	185	490	23	21.30435	0.046939
Maron	225	318	10	31.8	0.031447
Newman	185	517	5	103.4	0.009671
Stallings	225	374	8	46.75	0.02139
O'niel	200	482	34	14.17647	0.070539
Goldscmidt	220	603	31	19.45161	0.05141
Edman	180	641	11	58.27273	0.017161
Rojas	200	484	11	44	0.022727
Kelley	210	304	13	23.38462	0.042763
Versho	205	284	11	25.81818	0.038732
McHahon	219	528	23	22.95652	0.043561
Tapia	175	487	6	81.16667	0.01232
Rogers	204	387	15	25.8	0.03876
Bellinger	203	315	10	31.5	0.031746
Turner	185	595	28	21.25	0.047059
Pollock	210	384	21	18.28571	0.054688

Machado	218	564	28	20.14286	0.049645
Hosmer	226	509	12	42.41667	0.023576
Grisham	224	462	15	30.8	0.032468
Ruf	250	262	16	16.375	0.061069
Casali	220	200	5	40	0.025
Yastrzemski	180	468	25	18.72	0.053419

Regression Equation

AB/Hr = 98.1 - 0.3192 Weight

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	98.1	17.6	5.57	0.000	
Weight	-0.3192	0.0832	-3.84	0.000	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
17.9539	14.63%	13.63%	10.35%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	4749	4749.1	14.73	0.000
Weight	1	4749	4749.1	14.73	0.000
Error	86	27721	322.3		
Lack-of-Fit	35	11493	328.4	1.03	0.452
Pure Error	51	16228	318.2		
Total	87	32470			

Fits and Diagnostics for Unusual Observations

Obs	AB/Hr	Fit	Resid	Std Resid	
1	90.67	29.45	61.22	3.43	R
7	14.10	8.06	6.04	0.36	X
15	111.00	42.22	68.78	3.91	R
19	13.93	13.49	0.45	0.03	X
29	15.67	11.25	4.41	0.26	X
69	103.40	39.02	64.38	3.63	R
78	81.17	42.22	38.95	2.21	R

R Large residual
X Unusual X

Jacob Strong

Professor Caso

Advanced Placement Statistics

November 19, 2021

Analysis of Data

The home run production of a baseball player is an extremely important aspect, and many teams build around a player who succeeds at this. Just like in other sports, a team needs to be able to find prospects in order to create a competitive roster, thus the ability to assume how a player will do, deems a massive advantage. Notable names, like that of Paul DePodesta, changed how teams look at players, and he used statistics to determine who his team should invest in. The real question is, however, what characteristics make a good home run hitter; this study hopes to do exactly that. In this study, professional baseball players will be observed based on their weights and their at-bats per home run, in order to see if there is any valuable association between the two.

In hopes to see if there is an association between a player's weight and their home run production, this study looked at professional baseball players, in order to satisfy the population of interest. It must be said that any relationship cannot be generalized to every baseball player, for the sample taken was of only professionals. A sample of ninety ballplayers was taken, all of which were from the MLB, meaning that all the players were given relatively similar options to equipment and other extraneous factors. A form of stratified data collection took place, for every team was considered a stratum, and 3 players from each stratum were selected by giving each

player a number and randomly selecting a number. Players who didn't have at least two hundred at-bats were excluded from data collection, for if not, there would be players with abnormally high production due to fortune rather. The selected players were then listed, and their data was collected regarding their weights, at-bats, and home runs. All this data came from the Entertainment and Sports Programming Network (ESPN), for they are a massive organization and have uniform requirements for every one of the stats collected. From there, the at-bats per home run function was calculated (the home runs per at-bat was also calculated in case one found that more valuable), and each player had satisfactory data. It needs to be further stated that the population of interest was not baseball players in general, and instead professionals, thus one cannot stretch information from this study across players outside the MLB.

After all the data was collected and organized, analysis was completed in order to satisfy the intentions of the study. The original data had a linear regression line of AB/Hr = 133.2 - .4647(Weight), and further, that the value for r was -.283. This means that there is very little correlation between the two variables of weight and home run production, however there were obvious outliers in the data. From there, the outliers were removed and analysis was done again (this is the data provided above). This time around, the regression line was described as AB/Hr = 98.07 - .3192(weight), with an r value of -.382. The relationship found earlier still stands, and to do further analysis, the r^2 was calculated, and was valued at .146. This numerical description shows that there is around a 14% variance in home run production that can be attributed to the linear relationship between weight and home run production. In simple terms, there is a very dim relationship, and it is one that seems more extraneous than confounding. To check if the data could at all be transformed, a regression plot was created, and no pattern was found, meaning there was no transforming to be done. Finally, a histogram of weight was created to see if there

was any overwhelming value of weight present that may be skewing the data. However, the histogram was somewhat normal, and didn't give notice to any unfair distribution. One can, however, observe that there are several unusual observations, however they are still necessary to be kept in analysis, for they do present information that cannot be tossed out. After the study was completed, one can observe that the goal was accomplished, and that very little relationship was found between weight and home run production. This does not mean that there is no relationship, however it is minimal and ought not to be deemed as an important characteristic.

"MLB - Major League Baseball Teams, Scores, Stats, News, Standings, Rumors." *ESPN*, ESPN Internet Ventures, https://www.espn.com/mlb/.