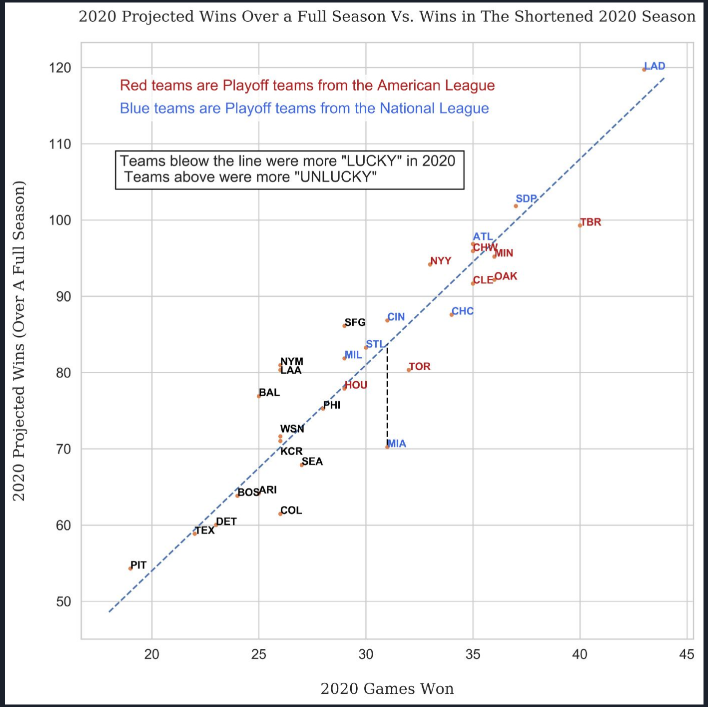
Regression in Baseball

Aidan Resnick

What do you think regression means?

Regression Definition

- Regression - a return to a former or less developed state
- Regression goes both ways
- Two Dimensions of Regression
    - Conceptual
        - Noise Regression
        - Regression to the Mean
    - Computational
        - Simple Linear Regression
        - Multiple Linear Regression
        - Logistic Regression



Noise Regression

- What do you see?
  - https://www.youtube.com/watch?v=KnuImyEvbPY (Start at 7:53)

Truth About Bregman's Performance

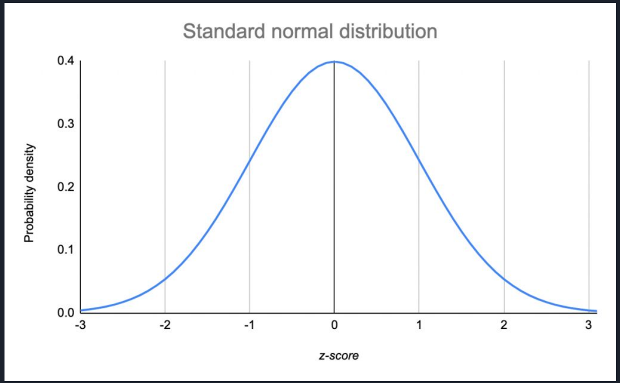
- Per Statcast, Bregman's batted ball had a 79% chance of falling for a hit.
  - However, Bregman received no credit for a hit on this play.
- The disparity between batting average and expected batting average can be attributed primarily to noise.

						Batting Avg			Slugging			Quality of Contact + K + BB		
Rk.	Player	Team	PA	Pos	BIP	BA	xBA	Diff	SLG	xSLG	Diff	wOBA	xwOBA	Diff
1	Goldschmidt, Paul		651	1B	424	.317	.261	0.056	.578	.482	0.096	.419	.367	0.052
2	Bogaerts, Xander		631	SS	446	.307	.259	0.048	.456	.383	0.073	.363	.323	0.040
3	McNeil, Jeff		589	2B	477	.326	.280	0.046	.454	.389	0.065	.365	.323	0.042
4	Giménez, Andrés		557	2B	386	.297	.257	0.040	.466	.400	0.066	.364	.326	0.038
5	González, Luis		350	RF	241	.254	.215	0.039	.360	.306	0.054	.302	.268	0.034
6	Contreras, William		376	C	232	.278	.243	0.035	.506	.479	0.027	.370	.347	0.023
7	Machado, Manny		644	3B	447	.298	.264	0.034	.531	.447	0.084	.382	.338	0.044
8	McCarthy, Jake		354	RF	249	.283	.249	0.034	.427	.357	0.070	.337	.298	0.039
9	Happ, Ian		641	LF	428	.271	.239	0.032	.440	.379	0.061	.339	.306	0.033
10	Taylor, Michael A.		456	CF	310	.254	.223	0.031	.357	.358	-0.001	.297	.289	0.008

						Batting Avg			Slugging			Quality of Contact + K + BB		
Rk.	Player	Team	PA	Pos	BIP	BA	xBA	Diff	SLG	xSLG	Diff	wOBA	xwOBA	Diff
1	Santana, Carlos		506	1B	345	.202	.253	-0.051	.376	.438	-0.062	.308	.352	-0.044
2	Toro, Abraham		352	2B	263	.185	.226	-0.041	.324	.372	-0.048	.246	.284	-0.038
3	Kepler, Max		446	RF	326	.227	.266	-0.039	.348	.412	-0.064	.298	.338	-0.040
4	Seager, Corey		663	SS	495	.245	.283	-0.038	.455	.510	-0.055	.331	.372	-0.041
5	De La Cruz, Bryan		355	CF	243	.252	.287	-0.035	.432	.498	-0.066	.313	.355	-0.042
6	Vierling, Matt		357	CF	261	.246	.279	-0.033	.351	.408	-0.057	.285	.327	-0.042
7	Tellez, Rowdy		599	1B	411	.219	.252	-0.033	.461	.479	-0.018	.327	.349	-0.022
8	Ozuna, Marcell		507	DH	352	.226	.256	-0.030	.413	.478	-0.065	.298	.337	-0.039
9	Winker, Jesse		547	LF	356	.219	.249	-0.030	.344	.403	-0.059	.313	.345	-0.032
10	Stanton, Giancarlo		452	DH	264	.211	.240	-0.029	.462	.477	-0.015	.327	.351	-0.024

General Truths

- In sports, there will almost always be external factors that affect the outcome of a player's performance.
  - In baseball, as previously displayed, a hitter could be robbed of a hit by a defender.
    - Similarly, a hitter could be given a hit due to a poor play from a defender.
  - In hockey, a skater could be robbed of a goal by the goaltender.
    - Similarly, a skater could be given a goal due to a poor play from the goaltender.
  - In football and basketball as well, the defense could be the reason why an offensive player did not score or the reason why an offensive player did score.
  - These examples could applied to many different sports
- In player analysis, it is always essential to expect noise to regress to zero, also known as its mean…



$$Y_i = f(X_i, \beta) + e_i$$

Y_i = dependent variable

f = function

X_i = independent variable

β = unknown parameters

e_i = error terms

Regression to the Mean

- Consider this information about Zack Greinke
- In 2015, as a member of the Los Angeles Dodgers, Zack Greinke posted an ERA of 1.66.
    - That is the lowest single-season ERA of the decade.
- He threw a 45.2 inning scoreless streak, fourth-longest in the expansion era.
- Runner-up in the Cy Young award race
- Earned him a $206.5 million contract

More information about Greinke

- In 2015, Greinke posted a 3.22 xFIP.
    - xFIP estimates ERA when accounting for random factors such as fielding, order of events, ballpark.
- Whereas his 1.66 ERA ranked first of the decade, a 3.22 xFIP is much more representative of an average top starter.
- So using this information, what would you expect Greinke's ERA to be in 2016?

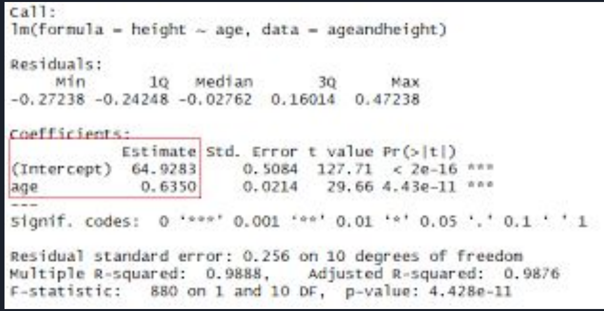
Even more information about Greinke

- In 2016, Greinke’s ERA was 4.37, increasing by 2.71 from the prior season.
  - To many, his first year on the Arizona Diamondbacks was seen as a colossal disappointment.
- But his xFIP was 3.98.
- So was Greinke truly *that* much worse in 2016 than in 2015?
  - If so, why do you think so?
  - If not, then why did his ERA increase by such a significant amount?

ERA Predictions for 2017?

Season	Team	W	L	SV	G	GS	IP	K/9	BB/9	HR/9	BABIP	LOB%	GB%	HR/FB	EV	ERA	FIP	xFIP	WAR
2004	Royals	8	11	0	24	24	145.0	6.21	1.61	1.61	.267	80.4%	34.6%	13.2%		3.97	4.70	4.30	1.8
2005	Royals	5	17	0	33	33	183.0	5.61	2.61	1.13	.335	65.2%	39.2%	9.6%		5.80	4.49	4.66	2.0
2006	Royals	1	0	0	3	0	6.1	7.11	4.26	1.42	.316	81.4%	35.0%	16.7%		4.26	5.04	4.32	0.0
2007	Royals	7	7	1	52	14	122.0	7.82	2.66	0.89	.314	75.6%	32.1%	7.4%		3.69	3.74	4.14	2.1
2008	Royals	13	10	0	32	32	202.1	8.14	2.49	0.93	.308	75.2%	42.7%	9.1%		3.47	3.56	3.71	4.2
2009	Royals	16	8	0	33	33	229.1	9.50	2.00	0.43	.303	79.3%	40.0%	4.5%		2.16	2.33	3.09	8.7
2010	Royals	10	14	0	33	33	220.0	7.40	2.25	0.74	.305	65.3%	46.0%	7.5%		4.17	3.34	3.60	4.9
2011	Brewers	16	6	0	28	28	171.2	10.54	2.36	1.00	.318	69.8%	47.3%	13.6%		3.83	2.98	2.56	3.3
2012	2 Teams	15	5	0	34	34	212.1	8.48	2.29	0.76	.306	74.5%	49.2%	10.2%		3.48	3.10	3.22	4.8
2013	Dodgers	15	4	0	28	28	177.2	7.50	2.33	0.66	.276	80.8%	45.6%	8.6%		2.63	3.23	3.45	3.4
2014	Dodgers	17	8	0	32	32	202.1	9.21	1.91	0.85	.311	79.7%	48.7%	11.9%		2.71	2.97	2.72	4.5
2015	Dodgers	19	3	0	32	32	222.2	8.08	1.62	0.57	.229	86.5%	48.0%	7.3%	87.4	1.66	2.76	3.22	5.3
2016	Diamondbacks	13	7	0	26	26	158.2	7.60	2.33	1.30	.294	71.8%	45.9%	13.9%	87.9	4.37	4.12	3.98	2.3

Computational Regression

- Over the course of our club meetings, we will discuss several types of regression
- Most common:
    - Simple Linear Regression
        - Predicting outcome Y with predictor X
    - Multiple Linear Regression
        - Predicting outcome Y with predictor $X_1, X_2, \ldots, X_n$
    - Logistic Regression
        - Predicting binary outcome Y with predictors X_n
- Regression methods in R and Python show relationship strength (correlation)
    - 

```
Call:
lm(formula = height ~ age, data = ageandheight)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27238 -0.24248 -0.02762  0.16014  0.47238

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  64.9283     0.5084  127.71  < 2e-16 ***
age           0.6350     0.0214   29.66 4.43e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.256 on 10 degrees of freedom
Multiple R-squared:  0.9888,	Adjusted R-squared:  0.9876
F-statistic:   880 on 1 and 10 DF,  p-value: 4.428e-11
```

Questions?