

# Statistical Analysis of the Chicago Cubs

---

Johnna Anderson

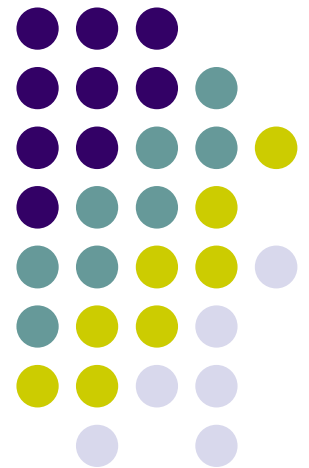
Scott Diehl

Jason Hatfield

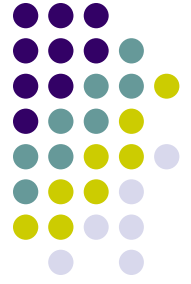
Nikita Tuzov

Shih-yi Wang

Prasanth Karumanchi



# Background



- Statistics in Sports
- Moneyball – Michael Lewis
- Mavericks and IU professor Wayne Winston



# Objective

- Predict *probabilities* and *means* of runs scored by Cubs using the following as predictor variables:
  - Walks
  - Hits
  - Home runs
  - Stolen Bases
  - Runners Left-on-Base
  - Strikeouts
  - Left/right Handedness of the opponents starting pitchers
  - Home or away game

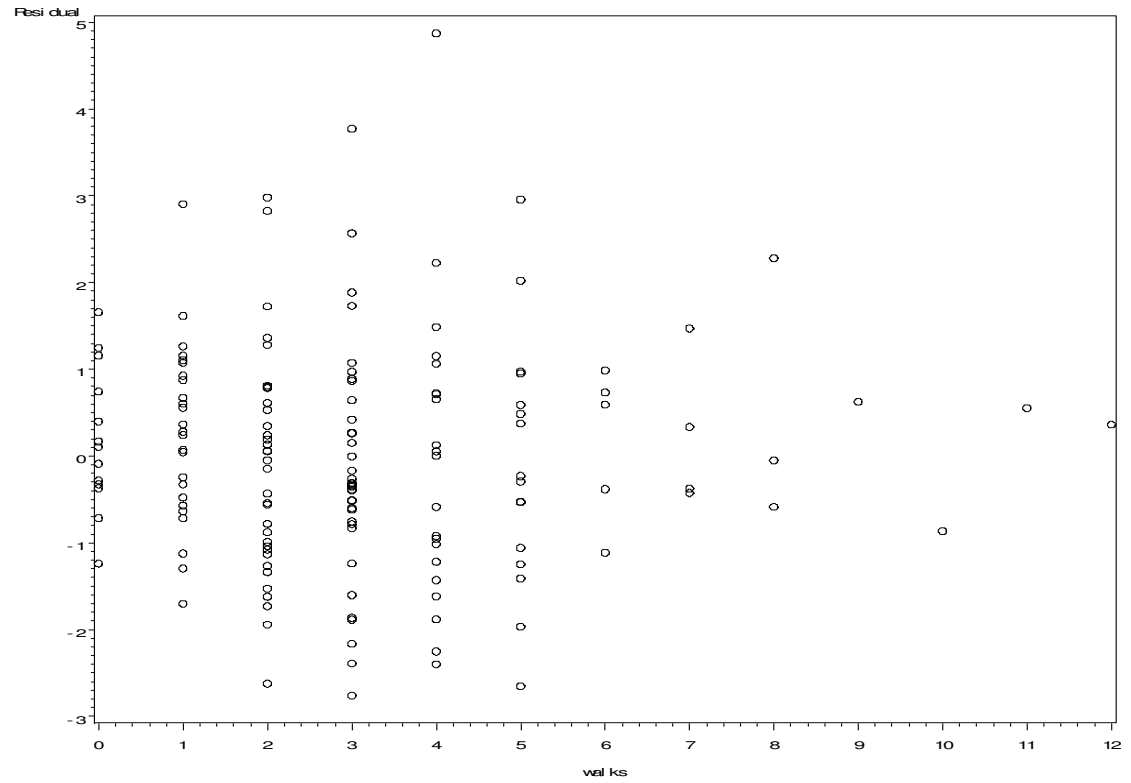


# Regression Models

- The Multiple Linear Regression Model
  - Potential shortcomings:
    - Model assumptions invalid (Non constant variance? Non normality?)
    - Unrealistic (negative) prediction of a nonnegative quantity (score in a baseball game.)
    - Estimating probabilities difficult (gives mean of a continuous distribution; our response variable is discrete.)

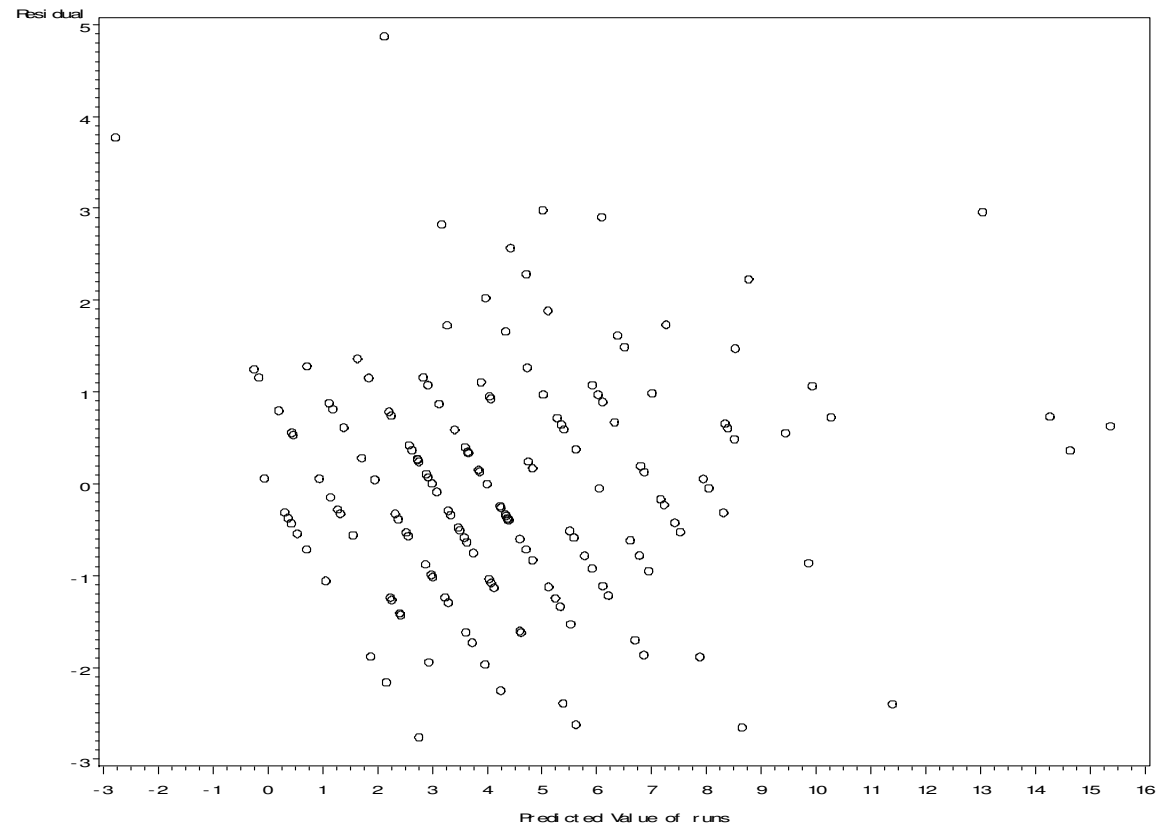


- Non-constant Variance?





- Negative prediction of runs





# Regression Models

- Weighted Regression Model: no visible change in estimates.
- Normality assumption may also be an issue:
  - Normality of Variance Tests (Results: it's not an issue.)

*Test for Normality (Weighted Least Square)*

*Tests for Normality*

<i>Test</i>	<i>--Statistic--</i>	<i>-----p Value-----</i>
<i>Shapiro-Wilk</i>	<i>W 0.979426</i>	<i>Pr &lt; W 0.0163</i>
<i>Kolmogorov-Smirnov</i>	<i>D 0.050446</i>	<i>Pr &gt; D &gt;0.1500</i>
<i>Cramer-von Mises</i>	<i>W-Sq 0.075892</i>	<i>Pr &gt; W-Sq 0.2375</i>
<i>Anderson-Darling</i>	<i>A-Sq 0.560663</i>	<i>Pr &gt; A-Sq 0.1484</i>



## Model Alternatives, contd.

- The unreduced, first-order MLR is not problematic; but our model can be improved.
- To obtain a more definitive model, we looked into:
  - Poisson Regression Model
  - Multiple Linear Regression Model with second-order terms

# Poisson Regression Model



- A Poisson model has features that are desirable for our situation:
  - It results in an estimated mass function rather than a density function. This makes prediction of probabilities easy for a discrete quantity like ours.
  - It models the natural log of the response variable,  $\ln(Y)$ , as a linear function of the coefficients, hence we get positive predicted means.



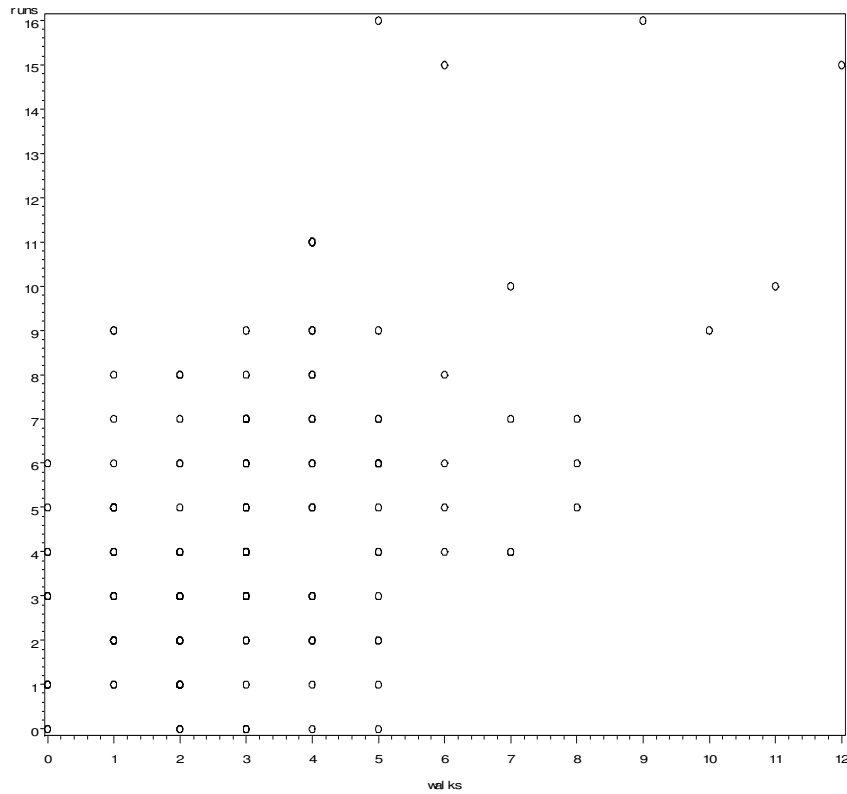
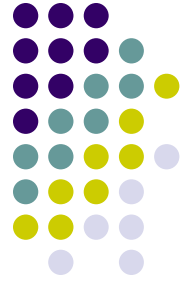
# Tests

Partial Deviance Test (after dropping non-significant variables):

$$\begin{aligned} & DEV(SB, HA, OppSP, SO | hits, walks, HR, LeftBase) \\ &= DEV(hits, walks, HR, LeftBase) - DEV(Full) \\ &= 116.9339 - 111.8516 \\ &= 5.0823 < \chi^2_{0.95;4} (= 9.4877) \end{aligned}$$

- Hence, we do not reject  $H_0$ . So, SB, HA, OppSP, SO are not found to be significant in the presence of hits, walks, HR and leftbase to predict the number of runs scored.

# Outliers



- A few data points suggest outlying X-values.

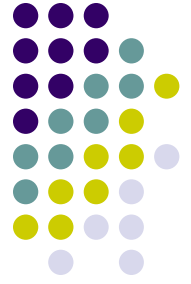


# Tests Contd.

- Hat Matrix Leverage to test for X-outliers:

Compared with  $2p/n = 2*5/162 = 0.0617$

<i>Obs</i>	<i>Residual</i>	<i>Hat Diag</i> <i>RStudent</i>	<i>Cov</i> <i>H</i>	<i>Ratio</i>	<i>DFFITS</i>
1	0.1889	0.1546	<b>0.1366*</b>	1.1950	0.0615
2	3.6754	3.0803	<b>0.1266*</b>	0.8799	1.1727
3	0.6065	0.4670	0.0233	1.0497	0.0721
4	0.4932	0.3805	0.0275	1.0567	0.0640
5	-0.7393	-0.5894	<b>0.0880*</b>	1.1196	-0.1831
6	-0.1850	-0.1418	0.0154	1.0480	-0.0178
7	1.8246	1.4116	0.0215	0.9902	0.2090
8	-0.5340	-0.4122	0.0284	1.0569	-0.0705
9	-0.9276	-0.7165	0.0276	1.0444	-0.1206
10	-1.4453	-1.1180	0.0259	1.0184	-0.1821
11	1.0942	0.8388	0.0114	1.0212	0.0902
12	0.2535	0.1976	0.0480	1.0832	0.0444
13	-0.3256	-0.2513	0.0286	1.0608	-0.0432
14	2.3434	1.8302	0.0316	0.9587	0.3307
15	0.3636	0.3025	<b>0.1639*</b>	1.2312	0.1339
16	0.3626	0.2936	<b>0.1173*</b>	1.1665	0.1070



## Tests Contd.

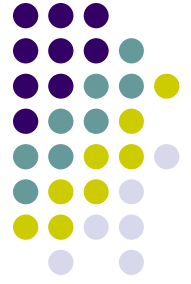
- When model is run without outlying observations, noticeably different estimates are obtained.
- One objective of the model is prediction of mean number of runs scored; we used playoff data for a comparison.
- We ran the comparison both with and without outlying  $X$ 's.

# Results



Runs	Prediction with Full Data	Prediction without Outliers
4	4.772165	5.401086
3	3.59089	4.054389
3	2.530963	1.997307
4	4.879291	5.441202
5	5.010321	5.359657
8	6.426307	7.656079
12	12.2828	18.72389
5	4.892483	5.387061
8	6.134938	8.987188
0	1.752775	1.325249
3	3.618284	3.214912
6	3.643337	3.57727

# Results



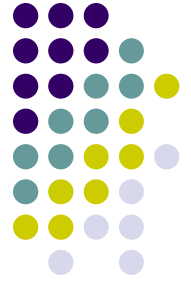
- Poisson Model (Without Outliers)
  - $E(\text{Runs}) = \exp(0.2114 + 0.1615 * \text{hits} + 0.2062 * \text{walks} + 0.1334 * \text{HR} - 0.1530 * \text{LeftBase})$
- Final Poisson Model (Outliers retained.)
  - $E(\text{Runs}) = \exp(0.4958 + 0.1271 * \text{hits} + 0.1208 * \text{walks} + 0.0803 * \text{HR} - 0.1032 * \text{LeftBase})$

# Multiple Linear Regression Model With Second Order Terms



- Advantages of the model
  - Plenty of  $Df$ 's; can afford to include higher terms.  
(No over fitting.)
  - Reduces potential Multicollinearity.
  - No negative predictions were seen.

# MLR w/Second Order Terms continued...



- Four squared predictors (hits, walks, stolen bases, Leftbase) and all six of their interactions were included in the model. Based mostly on the visual analysis of all graphs of response vs single predictor.
- **C<sub>p</sub>** criterion was used to select the best set of predictors. As a result, “only” 11 variables were retained.
- Two predictors – Home/Away and Strikeouts – were dropped in first order.



# More MLR w/Second Order

## Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	1678.70283	152.60935	86.35	<.0001
Error	145	256.26078	1.76732		
Corrected Total	156	1934.96361			
Root MSE	1.32940			R-Square	0.8676
Dependent Mean	3.04714			Adj R-Sq	0.8575
Coeff Var	43.62797				

## Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Tolerance
Intercept	1	0.28822	0.16446	1.75	0.0818	.
hits	1	0.39658	0.02290	17.32	<.0001	0.41303
walks	1	0.38765	0.02911	13.31	<.0001	0.61015
HR	1	0.15562	0.04949	3.14	0.0020	0.60166
SB	1	0.20123	0.09373	2.15	0.0335	0.83581
OppSP	1	-0.26993	0.12399	-2.18	0.0311	0.93554
LeftBase	1	-0.32982	0.02941	-11.22	<.0001	0.39176
cSO2	1	-0.00609	0.00232	-2.62	0.0097	0.90407
cLeftBase2	1	0.01051	0.00703	1.49	0.1373	0.46740
hitSO	1	-0.00534	0.00580	-0.92	0.3585	0.84029
hitLeft	1	-0.00384	0.00717	-0.54	0.5930	0.45986
walkSO	1	0.02817	0.00973	2.90	0.0044	0.81522

# Comparison of Predicted Values- Poisson vs. MLR



Playoff Game	Runs	Pred Runs MLR	Pred Run Poisson	Pred Runs, Poisson outliers deleted	MLR error sq	Poisson error sq	Poisson error sq outliers deleted
1	4	6.65	4.77	5.40	7.00	0.60	1.96
2	3	3.43	3.59	4.05	0.19	0.35	1.11
3	3	2.90	2.53	2.00	0.01	0.22	1.01
4	4	5.34	4.88	5.44	1.80	0.77	2.08
5	5	5.58	5.01	5.36	0.33	0.00	0.13
6	8	7.22	6.43	7.66	0.61	2.48	0.12
7	12	11.70	12.28	18.72	0.09	0.08	45.21
8	5	5.41	4.89	5.39	0.17	0.01	0.15
9	8	6.43	6.13	8.99	2.47	3.48	0.97
10	0	0.34	1.75	1.33	0.11	3.07	1.76
11	3	3.73	3.62	3.21	0.53	0.38	0.05
12	6	3.59	3.64	3.58	5.82	5.55	5.87
				<b>MSEs:</b>	<b>1.59</b>	<b>1.42</b>	<b>5.03</b>

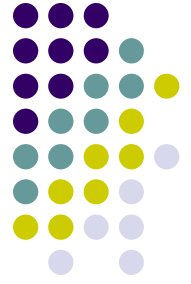


# Conclusions

- The prediction without using outliers was less accurate in our situation, since we had an outlier in the out-of-sample data.
- MLR with second order terms produced similar predictions to the Poisson model, but with slightly higher MSE and a more complicated model.
- Poisson model is more appropriate for answering questions concerning estimated probability, as we are modeling Poisson-distributed counts. For example, we can predict the probability of, say, 5 runs or more in a game, given counts of other quantities.



Questions ??



Thank you  
Go Cubs!