# Handedness Impact on Reaching First Base

Jingyuan Chen, Eric Gerber, Nathan Hankey, Qi Wang, Hao Zhang

*Department of Statistics, Purdue University.*
*250 N. University Street, West Lafayette, IN 47907-2066, USA*

May 12, 2014

**Abstract**

Baseball, it is said, is a game of inches. In Baseball, a left-handed batter starts out approximately 46.38 inches (nearly 4 feet) closer to first base than right-handed batters. If two players are of similar build, but start out in these different positions, does it make sense that the left-handed batter will have a better shot at reaching first base on a close play? Taking the simple statistic of singles, this study utilizes negative binomial regression to examine whether handedness affects a batters ability to get to first base, and thus if there is a significant advantage to being left-handed over right-handed.

*Keywords:* Poisson regression with exposure, negative binomial regression, MLB.

## 1. Introduction

### *1.1. Terminology*

This is a quick review of Baseball terminology. Those most germane to this study are presented in list form with accompanying explanations below:

- Batter: Offensive player who waits for ball to be thrown and attempts to hit said ball.
- Home Plate: Base where hitters stand on a given side of and wait to hit the ball.
- Batters Box: Area hitter is required to stand when hitting; one on either side of home plate.

- First Base: Located 90 feet from Home Plate; Hitters first destination after hitting ball.
- Single: Occurs when hitter successfully hits ball into field of play and reaches first base safely.
- Runs: The term for points scored by a team.

### *1.2. Background*

When introducing someone to the game of Baseball, a common way to start the explanation is with a description of the field dimensions, namely that the bases are all 90 feet apart and the pitching mound is 60 feet 6 inches from home plate. Then its explained that to score runs, the player batting wants to hit the ball thrown by the pitcher with the intent of making it on base and eventually going all the way around to score a run. The first stop in the path around the bases is first base, and thus reaching that base is necessary for any runs to be scored. Each batter is given the option, based on their preference, to hit right- or left-handed by standing in the appropriate batters box[1].

It is here that a possible disparity occurs. Assuming a point where the hitter stands in either batters box is even with the middle of home plate and equal distance from both sides of the batters box, a left-handed batter would start out approximately 46.38 inches closer to first base than a right-handed batter. In a game where plays, especially at first base, are decided by inches, this distance could possibly provide a significant advantage to left-handed batters when attempting to reach first base on a single, assuming that hitter is of the same build as a right-handed batter attempting to do the same.

### *1.3. Data set*

The dataset utilized in this study is comprised of information on the 458 offensive players (batters) with the most plate appearances over the course of the 2013 Major League Baseball season. The dataset also includes the batters Handedness (Left, Right or Both), their body mass index, $BMI = \frac{mass(kg)}{height(in)^2}$, and number of successes reaching first base (Singles). In addition, other variables of interest were included; a categorical version of BMI (cBMI) separating batters into Obese, Overweight, or Normal, batters team, and

[1]Figure 4 in Appendix

team Runs per Game (RpG). These were included given some concern of a potential covariance structure existing, i.e. a lack of independence due to a team a batter belongs to being better than another, which could affect the response. While there likely isnt a significant effect given that the baseball season is quite long: 162 games, it seemed prudent to consider. Collection of baseball data is such that there are no missing values, and outliers will be addressed. All data was taken from baseball-reference.com[2] , an online depository of all metrics used to measure on-field performance in the sport.

## 2. Methods

### *2.1. Poisson Regression*

The problem of interest for this study is to determine if a baseball batters handedness has an effect on the number of singles over a season, controlling for the physical size of a player (their build). Since singles, the response variable, are count data, a Poisson model seemed appropriate. Based on the distribution of singles from the dataset, this did not seem improper[3]. Also checked were the distribution of singles by handedness and it seems, before running regression, handedness is a good candidate for predicting number of singles since the mean appears to vary by the different levels[4].

Poisson regression also has the ability to account for opportunity. Major League Baseball batters do not achieve the same number of plate appearances throughout a season, thus a simple model without taking into account this exposure would not make sense. Poisson models can incorporate an offset, an exposure variable, in this case plate appearances (PA). Poisson models use the log as the canonical link function, resulting in the first model fit, our model of interest, taking the form:

$$\log \lambda_i = \log \textit{exposure}_i + X_i'\beta$$

Where $Y_i \sim \textit{Poisson}(\lambda_i)$, $\lambda_i$ is the number of singles, $\textit{exposure}_i$ is the number of plate appearances, and $X_i'\beta$ takes the form:

[2]Table 3: Excerpt in Appendix
[3]Figure 5 in Appendix
[4]Figure 6 in Appendix

$$X_i'\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Where $X_1$ is Handedness (0=Left, 1=Both, 2=Right), and $X_2$ is BMI. Initially the model fit treated $\log PA$ as an independent predictor, to determine the appropriateness of its use as an offset. It was found that the coefficient of $\log PA$ was close to 1 , meaning the offset is appropriate, since the model fit above is equivalent to using PA as a predictor with a coefficient of 1[5]. Next was to fit the model of interest:

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.5457     0.4480   1.218   0.2232
log(BMI)        -0.7319     0.1354  -5.405 6.47e-08 ***
Hand1           -3.0803     0.7799  -3.949 7.83e-05 ***
Hand2           -1.2842     0.5767  -2.227   0.0260 *
log(BMI):Hand1   0.9433     0.2355   4.006 6.18e-05 ***
log(BMI):Hand2   0.3984     0.1741   2.289   0.0221 *
---

Null deviance: 1014.03  on 457  degrees of freedom
Residual deviance:  966.65  on 452  degrees of freedom
AIC: 3605.7
```

Comparing the two models, the coefficients did not change drastically, and all the predictors are found to be significant. However, both the AIC and Residual deviance have risen, and there is evidence of overdispersion. Overdispersion occurs when the expected variance is larger than the expected mean. Overdispersion is estimated with a parameter ($\phi$) which is contrived from the $\chi^2$ Pearson lack of fit statistic. When this $\phi$ is greater than 1, there is evidence of overdispersion. For this model, $\phi = 2.1$.[6] A graph of the $\log E(Y)$ and $\log Var(Y)$ also shows that the variance is larger than the mean.[7] The presence of overdispersion (more fully discussed in the Model Diagnostics section) provides an incentive to investigate a negative binomial model as

[5]Output 1 in Appendix
[6]Output 2 in Appendix
[7]Figure 7 in Appendix

the next step, which allows for extra variation by treating the mean of the Poisson distribution as a random variable from the gamma distribution.

*2.2. Negative Binomial Regression*

Compared to the Poisson model, the assumptions for the negative binomial model are less intuitive. However, it allows for an extra parameter, which is able to control the overdispersion issue encountered in the Poisson setup. What follows briefly is a discussion of the model assumptions. Suppose that $Y|\lambda \sim Poisson(\lambda)$, $\lambda \sim \Gamma(k, \alpha)$, so that the joint distribution of Y and $\lambda$ is:

$$p(Y = y, \lambda) = \frac{\alpha^k}{\Gamma(k)y!}\lambda^{y+k-1}e^{-(\alpha+1)\lambda}$$

Then, by integrating $\lambda$ we get the marginal distribution of Y:

$$y = \frac{\alpha^k}{\Gamma(k)y!}\int_0^{+\infty} \lambda^{y+k-1}e^{-(\alpha+1)\lambda}d\lambda = \binom{y+k-1}{k-1}(\frac{\alpha}{\alpha+1})^k(\frac{1}{\alpha+1})^y$$

According to the marginal distribution of Y, this yields:

$$Y + k \sim NB(\frac{\alpha}{\alpha+1}, k)$$

And the expectation and variance of Y can then be calculated accordingly:

$$E\{Y\} = E\{Y + k\} - k = \frac{k}{\alpha} \overset{denote}{=} \mu$$

$$Var\{Y\} = \frac{k}{\alpha} + \frac{k}{\alpha^2} = \mu + \frac{\mu^2}{k}$$

Under these model assumptions, it must be determined whether it is appropriate to fix the coefficient of $\log PA$ to be 1 by considering it an offset; just as was done for the Poisson model. After fitting the negative binomial model without treating $\log PA$ as an offset, once again its found that an offset is appropriate[8]. The coefficient of $\log PA$ is very close to one, and the deviance test result,

$$G^2 = 481.84 - 455.02 = 26.82 > \chi^2_6(0.95) = 12.59159$$

Indicates that the other factors significantly affect the number of singles.

[8]Output 3 in Appendix

*2.3. Model Selection*

Next, several different candidate models are fit to determine which model will serve as the final model. After some initial exploratory model evaluation, the most important variables to include in our model selection process were identified. Below, each assuming a negative binomial distribution and treating plate appearances as an offset, are the eight models of interest.

1. $\log E\{Single_i\} = \beta_0 + \log PA + \beta_1 \log BMI + \beta_2 Hand_1 + \beta_3 Hand_2 + \beta_4 \log BMI \times Hand_1 + \beta_5 \log BMI \times Hand_2$
2. $\log E\{Single_i\} = \beta_0 + \log PA + \beta_1 \log BMI + \beta_2 Hand_1 + \beta_3 Hand_2$
3. $\log E\{Single_i\} = \beta_0 + \log PA + \beta_1 \log BMI + \beta_2 Hand_1 + \beta_3 Hand_2 + \beta_4 \log BMI \times Hand_1 + \beta_5 \log BMI \times Hand_2 + \beta_6 \log RpG$
4. $\log E\{Single_i\} = \beta_0 + \log PA + \beta_1 \log BMI + \beta_2 Hand_1 + \beta_3 Hand_2 + \beta_4 \log RpG$
5. $\log E\{Single_i\} = \beta_0 + \log PA + \sum_{i=1}^{3} \beta_i cBMI_i + \sum_{j=4}^{5} \beta_j Hand_j + \log RpG$
6. $\log E\{Single_i\} = \beta_0 + \log PA + \sum_{i=1}^{3} \beta_i cBMI_i + \sum_{j=4}^{5} \beta_j Hand_j + \sum_{i=1}^{3} \sum_{j=4}^{5} \beta_{ij} cBMI_i \times Hand_j + \log RpG$
7. $\log E\{Single_i\} = \beta_0 + \log PA + \sum_{i=1}^{3} \beta_i cBMI_i + \sum_{j=4}^{5} \beta_j Hand_j + \sum_{i=1}^{3} \sum_{j=4}^{5} \beta_{ij} cBMI_i \times Hand_j$
8. $\log E\{Single_i\} = \beta_0 + \log PA + \sum_{i=1}^{3} \beta_i cBMI_i + \sum_{j=4}^{5} \beta_j Hand_j$

The differences in the candidate models involved treating BMI as either continuous or categorical, including interaction between player build and handedness (BMI*Hand) or using an additive model, and lastly, including RpG (a gauge of team success) as a hedge against there being an effect of correlation between predictors, i.e. determining if the measurements possibly contain violations of the independence assumption.

Exploring a set of candidate models that represent data usually takes the form of a two-step process involving cross-validation with a training set to choose a model and then evaluating that models predictive ability with an evaluation set. The training set and evaluation set are each a randomly chosen subset of the full data set. However, this two-step process is a compromise most studies must adhere to due to a lack of data. Since the data set utilized in this study is relatively large, it is possible to perform a more rigorous three-step process, in which the data set is split into three subsets; a training set, a selection set, and an evaluation set.

## 3. Results

### *3.1. Prediction*

For this data, 45 observations were randomly sampled to serve as the evaluation set. Then, for the remaining 413 observations, 41 observations are sampled as the model selection set and the rest as the training set. The training set is used to fit the negative binomial models and the selection set is used to compare the difference between the predicted number of singles and the observed number.

After splitting the data in R, we construct each of our candidate models with the training set, and then use both RMSE (Residual Mean Square Error) and AIC to rank the candidates. While the AIC will not change for each model, the RMSE changes depending on the data used to fit the model. Instead of fitting each model based on a single training set and selection set, the process was conducted 10 times to provide an average RMSE for each model based on 10 different, randomly chosen training and selection sets. Below is a table of each model, their corresponding average RMSE and AIC.

Mean Square Error and AIC for each candidate model

| | model 1 | model 2 | model 3 | model 4 |
|---|---|---|---|---|
| mRMSE | 6078.90 | 6117.56 | 6076.09 | 6107.96 |
| AIC | 2772.74 | 2775.69 | 2773.10 | 2775.83 |
| | model 5 | model 6 | model 7 | model 8 |
| mRMSE | 6291.77 | 6216.15 | 6218.11 | 6303.45 |
| AIC | 2783.36 | 2781.28 | 2780.73 | 2783.06 |

While there is not an immense disparity across the board for our candidate models, based on both RMSE and AIC, we determined that Model 1 best represents the data and interpretation of story. Model 1 was fit on the evaluation set to determine predictive ability, and then fit on all the data to produce our final model.

Below is the model constructed out of the first two steps that will be used to evaluate predictive ability:

$$\log \lambda_i = offset(\log PA) + \log BMI * Hand$$

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.3625     0.8022   1.698 0.08943  .
log(BMI)        -0.9870     0.2428  -4.066 4.79e-05 ***
Hand1           -3.7355     1.2898  -2.896 0.00378  **
Hand2           -2.3307     1.0061  -2.317 0.02052  *
log(BMI):Hand1   1.1527     0.3894   2.960 0.00308  **
log(BMI):Hand2   0.7187     0.3038   2.366 0.01798  *
---

    Null deviance: 385.87  on 371  degrees of freedom
Residual deviance: 360.54  on 366  degrees of freedom
AIC: 2753.3
```

To help determine the models predictive utility, below is a plot of the predicted values of singles, $\exp \log \lambda_i$, and observed values for singles, using the evaluation set. With an $R^2 = 0.9101$, the plot shows that Model 1 fits well, and all the predictors appear significant in this situation. For the model fit on the full data set, the same plot of predicted vs. observed values is presented in the appendix [9].

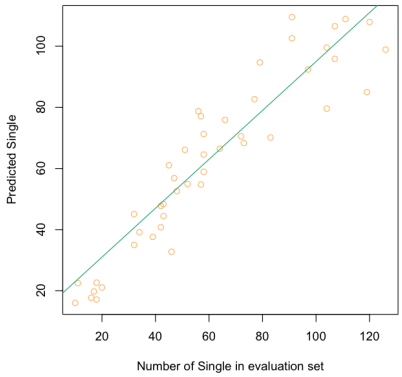
*3.2. Model Diagnostics*

Keeping the final model in mind, overdispersion in the original Poisson model was more fully investigated by examining the variance and mean of the response singles by each handedness. The variance within each level of handedness is significantly higher than the means within each level. These conditional means and variances further solidify the discovery of overdispersion and the necessity of a negative binomial model.

For the final model, residual plots were used to check the appropriateness scale for the fitted values. Linear predictors were plotted against both deviance residuals and response residuals.

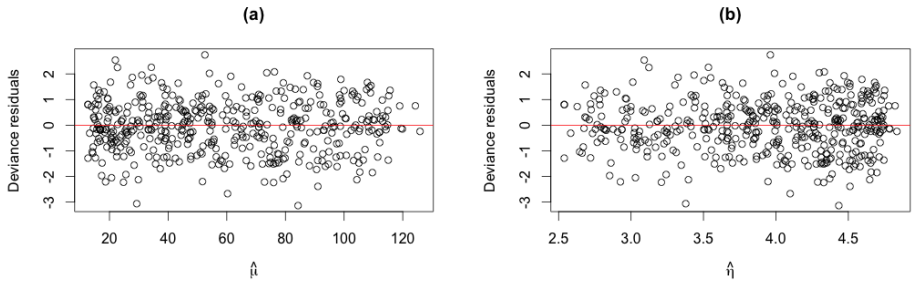
[9] Figure 8 in Appendix

Predicted number of Single VS observed number of Single



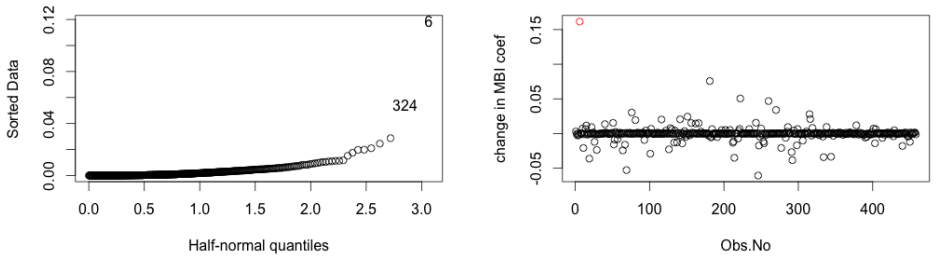

| | Hand | Mean | Variance |
|---|---|---|---|
| 1 | Left | 60.09 | 1105.774 |
| 2 | Both | 58.80 | 938.314 |
| 3 | Right | 58.75 | 1168.565 |

Mean-Variance

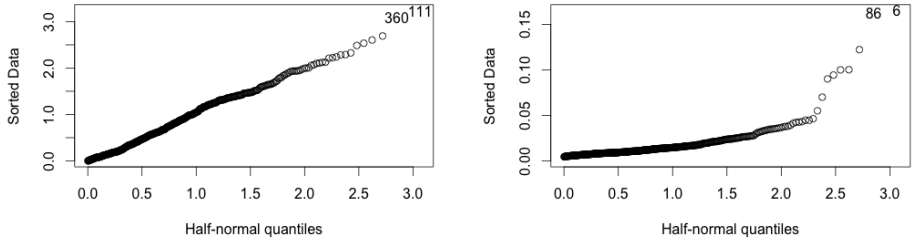



From the plots, it is clear there is a linear relationship between the predicted values and the residuals in each plot, which provide evidence of a good fit. By using the deviance residuals, the variance function has already been scaled out and as such there is expected to be constant variance in the plot, and that appears to be reflected in the output.

Half-normal plot of the Cook statistics left and an index plot of the change in the BMI coefficient right.




Half-normal plots of the jacknife residuals on the left and the leverages on the right and the residual plots (a) and (b) below.




In terms of checking the outliers, the raw material of the residuals, leverages and influence measures are utilized with half-normal plots. In the half-normal plot of the jackknife residuals there is no sign of outliers. However, by checking the leverage plot, case 6 and 86 appear to have significant leverage. From the original data these cases (Prince Fielder and Pablo Sandoval) have relatively higher BMI (38.35 and 34.2 respectively) when compared with others. These two players a considered among the heaviest in baseball, and Sandovals weight is even currently having a significant impact on his contract negotiations with his current team.

For further diagnostics, a half-normal plot of Cooks distance was produced. There is again some indication that Prince Fielder, a left-handed hitter, is significantly influential. This was the case across the board in terms of our diagnostics. Also examined was the change in the fitted coefficients against

individual observations. For illustration, change in BMI coefficient is shown and once again there is substantial change for Fielder. As the outlier he is, Fielder was excluded from the final model fit in 3.3.

*3.3. Final Model*

Constructed from fitting the final model to the complete dataset, the fit takes the form:

$$\log \lambda_i = 1.29 - 0.96 \log BMI - 3.87 Both - 1.98 Right$$
$$+1.19 \log BMIBoth + 0.61 \log BMIRight$$

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.2904     0.7253   1.779 0.075203 .
log(BMI)        -0.9617     0.2194  -4.383 1.17e-05 ***
Hand1           -3.8686     1.1900  -3.251 0.001150 **
Hand2           -1.9820     0.9095  -2.179 0.029309 *
log(BMI):Hand1   1.1868     0.3596   3.300 0.000967 ***
log(BMI):Hand2   0.6106     0.2746   2.223 0.026195 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

    Null deviance: 482.50  on 456  degrees of freedom
Residual deviance: 453.97  on 451  degrees of freedom
AIC: 3409.8
```

Model 1 maintains BMI as the continuous predictor it truly is, and keeps the interaction between BMI and our main predictor of interest, handedness. From this model, the intuition about left-handed batters seems to ring true; that being right-handed lowers the number of expected singles. Interestingly, being a batter that switches around results in an even worse expected performance. It is also found that a higher BMI decreases the expected counts as well, another result that seems to make sense; a heavier batter will not be as fast, and will thus reach first base less times than a more fleet of foot batter.

## 4. Discussion

The interpretation of the final model seems to prove some intuition about the problem addressed. Left-handed batters start nearly 4 feet closer to first base on average than right-handed hitters, and this model shows that that does appear to provide an advantage. This is taking into account the physical build of a player, as slower and heavier batters have a distinct disadvantage. Further, there seems to be some interaction between BMI and handedness, as certain body types appear to prefer a certain handedness. Left-handed batters and batters who hit on both sides generally appear to have smaller BMIs than right-handed batters[10]. The model does a good job of describing these characteristics, and with all predictors significant, and high correlation between predicted and observed values, it appears our model is more than adequate for not only showing that being left-handed provides a clear advantage when it comes to reaching first base, but there is potential to predict a batters number of singles based simply on their handedness and BMI.

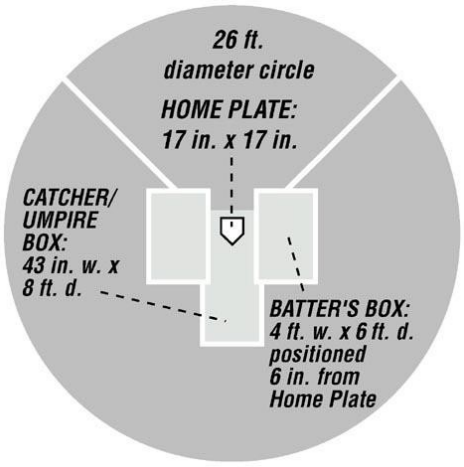
Practically, this is not an earth-shattering revelation, and the benefits are minor in the grand scheme of the sport. However, it provides evidence that when training someone to hit a baseball, it would be worthwhile for them to attempt to train as a left-handed batter to offer an extra advantage down the line.

[10]Figure 9 in Appendix

## 5. Appendix

[1]

**Figure 4**

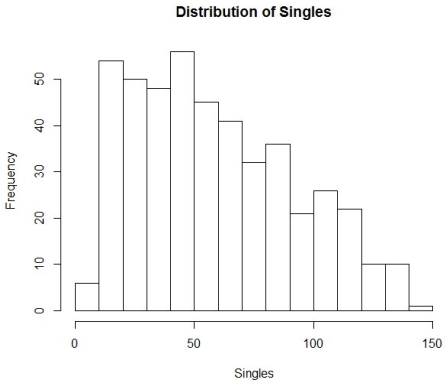



[2]

| | Single | PA | Age | Hand | BMI | Tm | cBMI | RpG |
|---|---|---|---|---|---|---|---|---|
| 1 | 120 | 726 | 29 | 0 | 28.24 | CIN | OverWeight | 4.48 |
| 2 | 140 | 724 | 29 | 2 | 25.09 | BOS | OverWeight | 5.27 |
| 3 | 126 | 717 | 27 | 0 | 26.87 | STL | OverWeight | 3.95 |
| 4 | 115 | 716 | 21 | 2 | 29.53 | LAA | OverWeight | 3.77 |
| 5 | 105 | 712 | 30 | 0 | 28.59 | CIN | OverWeight | 4.60 |
| 6 | 113 | 712 | 29 | 0 | 38.35 | DET | Obese | 4.48 |
| 7 | 107 | 710 | 25 | 2 | 30.92 | ARI | Obese | 4.00 |
| 8 | 121 | 710 | 20 | 2 | 23.11 | BAL | Normal | 3.69 |
| 9 | 117 | 705 | 23 | 2 | 25.77 | CHC | OverWeight | 3.82 |
| 10 | 115 | 700 | 29 | 0 | 29.02 | KCR | OverWeight | 4.91 |

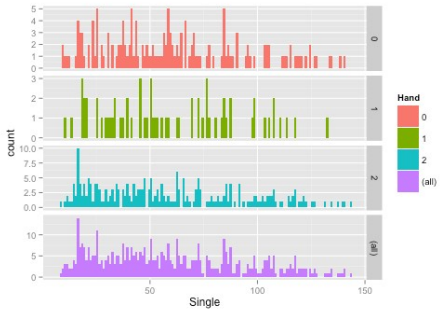
[3]

**Figure 5**




[4]

**Figure 6**




[5]

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.18147    0.45209   0.401  0.68813
log(PA)       1.09101    0.01361  80.185  < 2e-16 ***
log(BMI)     -0.78945    0.13591  -5.808 6.30e-09 ***
Hand1        -3.09464    0.78207  -3.957 7.59e-05 ***
Hand2        -1.58011    0.57892  -2.729  0.00635 **
```

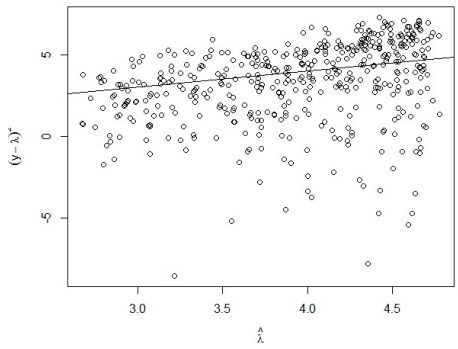
```
log(BMI):Hand1  0.94899    0.23614   4.019 5.85e-05 ***
log(BMI):Hand2  0.48871    0.17473   2.797  0.00516 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 8728.19  on 457  degrees of freedom
Residual deviance:  921.18  on 451  degrees of freedom
AIC: 3562.2
```

[6]

```
> sum(residuals(fit2, type="pearson")^2) / fit2$df.res
[1] 2.100066
```

[7]

**Figure 7**



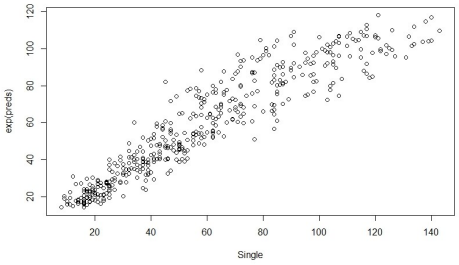

[8]

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.22197    0.67043   0.331  0.74058
log(PA)        1.08439    0.01831  59.228  < 2e-16 ***
log(BMI)      -0.82508    0.19858  -4.155 3.26e-05 ***
```

```
Hand1          -3.25358    1.13621  -2.864  0.00419 **
Hand2          -1.60260    0.85074  -1.884  0.05960 .
log(RpG)     0.08045    0.08718   0.923  0.35607
log(BMI):Hand1  1.00101    0.34326   2.916  0.00354 **
log(BMI):Hand2  0.49605    0.25669   1.932  0.05330 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 4475.7  on 457  degrees of freedom
Residual deviance:  451.4  on 450  degrees of freedom

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.45779    0.68060   0.673  0.501187
log(BMI)       -0.76992    0.20180  -3.815  0.000136 ***
Hand1          -3.22593    1.15663  -2.789  0.005286 **
Hand2          -1.35765    0.86456  -1.570  0.116337
log(RpG)     0.14220    0.08782   1.619  0.105397
log(BMI):Hand1  0.99034    0.34944   2.834  0.004595 **
log(BMI):Hand2  0.42085    0.26084   1.613  0.106654
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 481.84  on 457  degrees of freedom
Residual deviance:  455.02  on 451  degrees of freedom
```

[9]

**Figure 8**




[10] http://www.cbssports.com/mlb/writer/jon-heyman/24539024/panda-seeking-100m-plus-is-testing-giants-record-for-keeping-stars

[11]

**Figure 9**