



Baseball Case Study Analysis

Olusegun Omotunde

Motivation



- Baseball is the 3rd most popular sport in America with an average audience of 68.48 million according to Wikipedia.

Objective

- To determine the most significant variables on the salary
- To predict the salary of baseball players

Data


- Dataset of career stats of major league players.
- There are 320 observations and 20 variables in the dataset
- Mostly we have three major data types in the data (int, character and number), looking at the values of some observations to the right.
- The Salary column has 59 observations currently missing
- We therefore find ways to deal with the missing values

```
str(df0)
```

```
## 'data.frame':   322 obs. of  20 variables:
## $ AtBat      : int  293 315 479 496 321 594 185 298 323 401 ...
## $ Hits       : int  66 81 130 141 87 169 37 73 81 92 ...
## $ HmRun      : int  1 7 18 20 10 4 1 0 6 17 ...
## $ Runs       : int  30 24 66 65 39 74 23 24 26 49 ...
## $ RBI        : int  29 38 72 78 42 51 8 24 32 66 ...
## $ Walks      : int  14 39 76 37 30 35 21 7 8 65 ...
## $ Years      : int  1 14 3 11 2 11 2 3 2 13 ...
## $ CAtBat     : int  293 3449 1624 5628 396 4408 214 509 341 5206 ...
## $ CHits      : int  66 835 457 1575 101 1133 42 108 86 1332 ...
## $ CHmRun     : int  1 69 63 225 12 19 1 0 6 253 ...
## $ CRuns      : int  30 321 224 828 48 501 30 41 32 784 ...
## $ CRBI       : int  29 414 266 838 46 336 9 37 34 890 ...
## $ CWalks     : int  14 375 263 354 33 194 24 12 8 866 ...
## $ League     : chr  "A" "N" "A" "N" ...
## $ Division   : chr  "E" "W" "W" "E" ...
## $ PutOuts    : int  446 632 880 200 805 282 76 121 143 0 ...
## $ Assists    : int  33 43 82 11 40 421 127 283 290 0 ...
## $ Errors     : int  20 10 14 3 4 25 7 9 19 0 ...
## $ Salary     : num  NA 475 480 500 91.5 750 70 100 75 1100 ...
## $ NewLeague  : chr  "A" "N" "A" "N" ...
```

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
AtBat	0	1.00	380.93	153.40	16.0	255.25	379.5	512.00	687	
Hits	0	1.00	101.02	46.45	1.0	64.00	96.0	137.00	238	
HmRun	0	1.00	10.77	8.71	0.0	4.00	8.0	16.00	40	
Runs	0	1.00	50.91	26.02	0.0	30.25	48.0	69.00	130	
RBI	0	1.00	48.03	26.17	0.0	28.00	44.0	64.75	121	
Walks	0	1.00	38.74	21.64	0.0	22.00	35.0	53.00	105	
Years	0	1.00	7.44	4.93	1.0	4.00	6.0	11.00	24	
CAtBat	0	1.00	2648.68	2324.21	19.0	816.75	1928.0	3924.25	14053	
CHits	0	1.00	717.57	654.47	4.0	209.00	508.0	1059.25	4256	
CHmRun	0	1.00	69.49	86.27	0.0	14.00	37.5	90.00	548	
CRuns	0	1.00	358.80	334.11	1.0	100.25	247.0	526.25	2165	
CRBI	0	1.00	330.12	333.22	0.0	88.75	220.5	426.25	1659	
CWalks	0	1.00	260.24	267.06	0.0	67.25	170.5	339.25	1566	
PutOuts	0	1.00	288.94	280.70	0.0	109.25	212.0	325.00	1378	
Assists	0	1.00	106.91	136.85	0.0	7.00	39.5	166.00	492	
Errors	0	1.00	8.04	6.37	0.0	3.00	6.0	11.00	32	
Salary	59	0.82	535.93	451.12	67.5	190.00	425.0	750.00	2460	

Missing value Exploration



- Multiple approaches were tried to deal with missing values, we settled with dropping NAs.
- We obtained better results with dropping NAs than median or mean computation.
- We fitted a full model and the Adjusted R-squared became 0.51.
- In the next steps we try to improve the adjusted R-squared using transformations and variable selection.

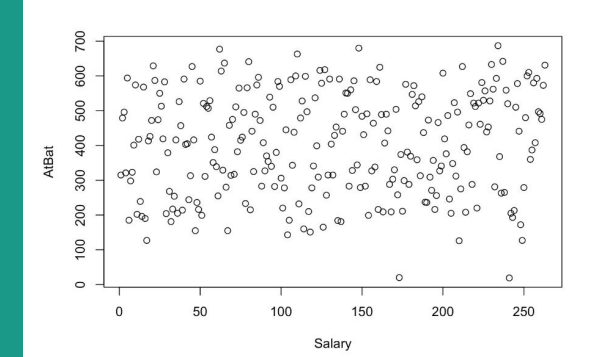
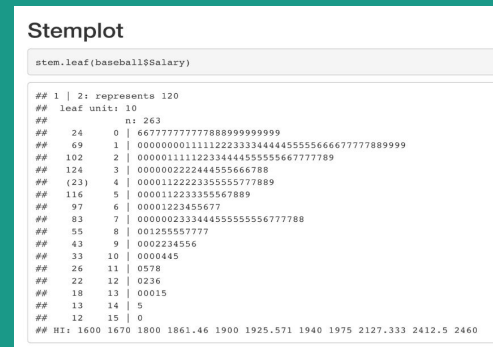
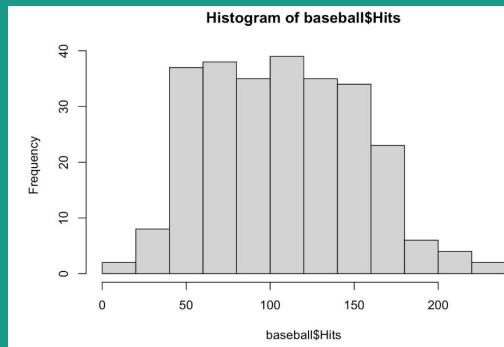
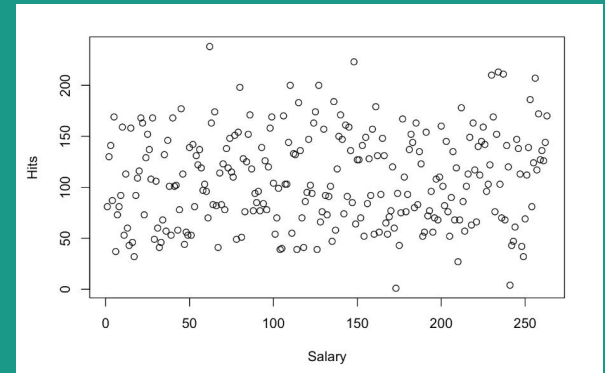
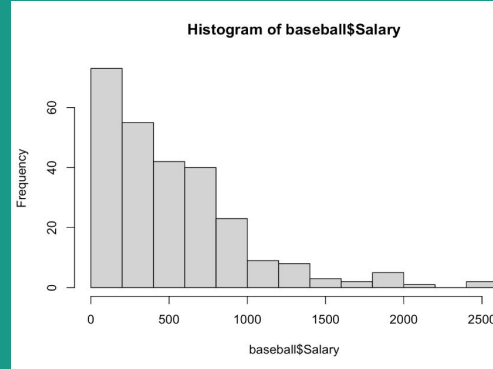
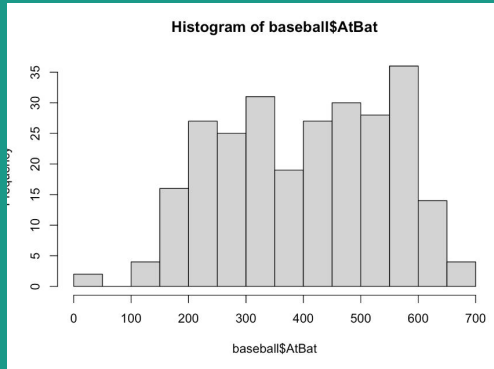
Response Variable Classification

For making a batch symmetric and for stabilizing spread across batches, we tried to reexpress the data.

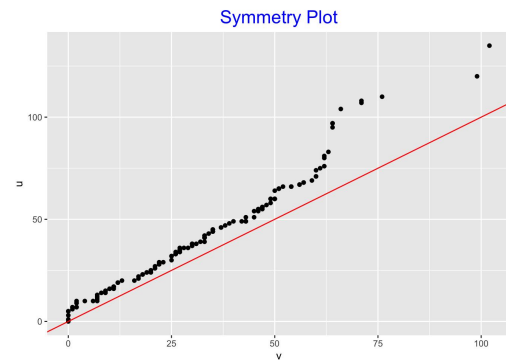
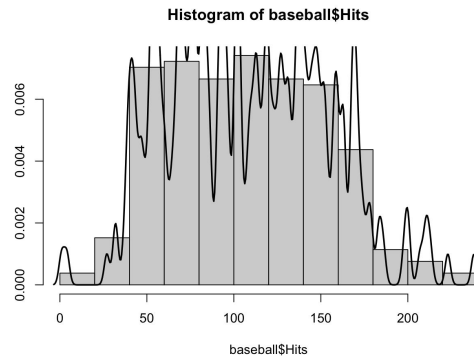
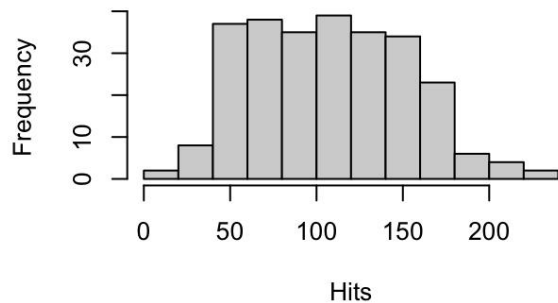
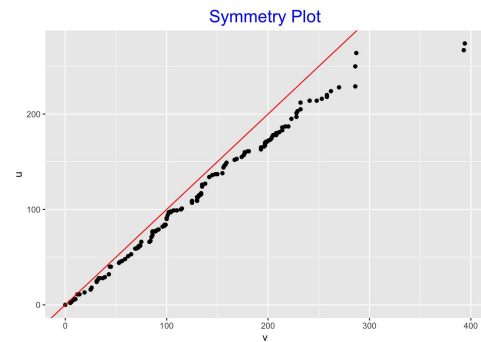
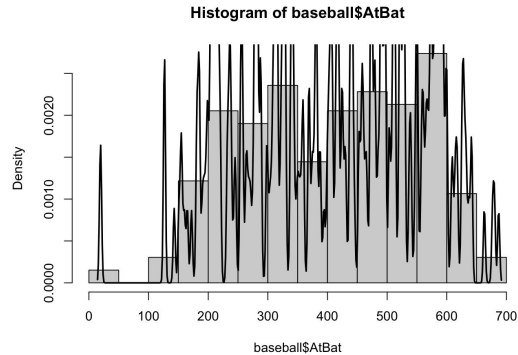
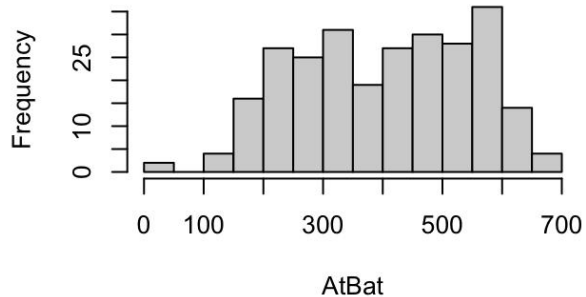
We used the following classification:

- Salary between 0 and 400 as 1
- Salary between 400 and 800 as 2
- Salary between 800 and 1200 as 3
- Salary between 1200 and 1600 as 4
- Salary between 1600 and 2000 as 5
- Salary between 2000 and 2500 as 6

Distribution of Important Variables



Two important variables: AtBat & Hits



Outliers

- There are no outliers in the data

```
d2 <- lval_plus(baseball, baseball$Hits)
filter(d2, OUT == TRUE)
```

```
## [1] AtBat Hits HmRun Runs RBI Walks Years
## [8] CatBat CHits CHmRun CRuns CRBI CWalks League
## [15] Division PutOuts Assists Errors Salary NewLeague Fence_LO
## [22] Fence_HI OUT
## <0 rows> (or 0-length row.names)
```

```
d3 <- lval_plus(baseball, baseball$AtBat)
filter(d3, OUT == TRUE)
```

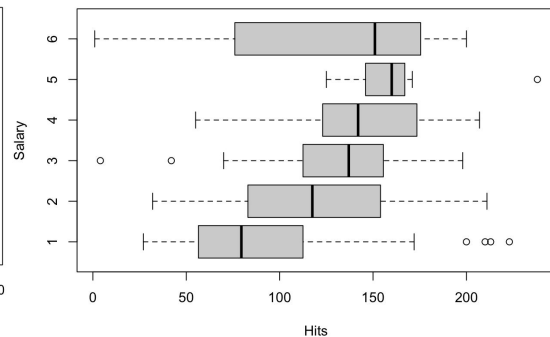
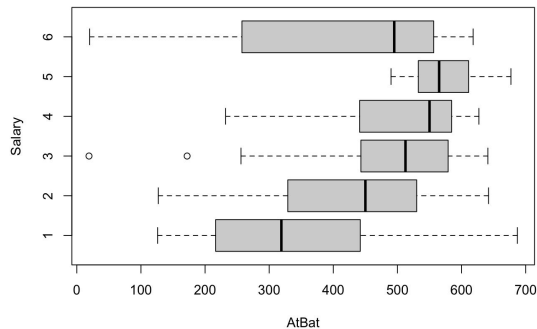
```
## [1] AtBat Hits HmRun Runs RBI Walks Years
## [8] CatBat CHits CHmRun CRuns CRBI CWalks League
## [15] Division PutOuts Assists Errors Salary NewLeague Fence_LO
## [22] Fence_HI OUT
## <0 rows> (or 0-length row.names)
```

```
baseball = baseball[which(baseball$AtBat>0),]
intersect(d2,d3)
```

```
## [1] AtBat Hits HmRun Runs RBI Walks Years
## [8] CatBat CHits CHmRun CRuns CRBI CWalks League
## [15] Division PutOuts Assists Errors Salary NewLeague Fence_LO
## [22] Fence_HI OUT
## <0 rows> (or 0-length row.names)
```

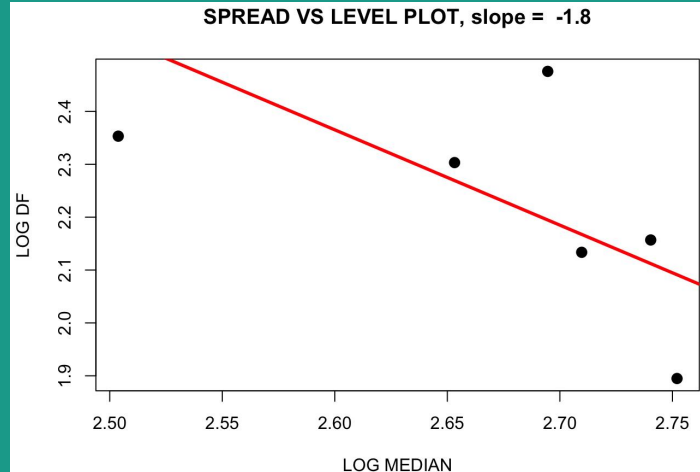
Boxplots: AtBat & Hits

- For different Salaries, we have different AtBat box plots and Hits box plots. They have different fourth spreads, which makes comparison difficult.

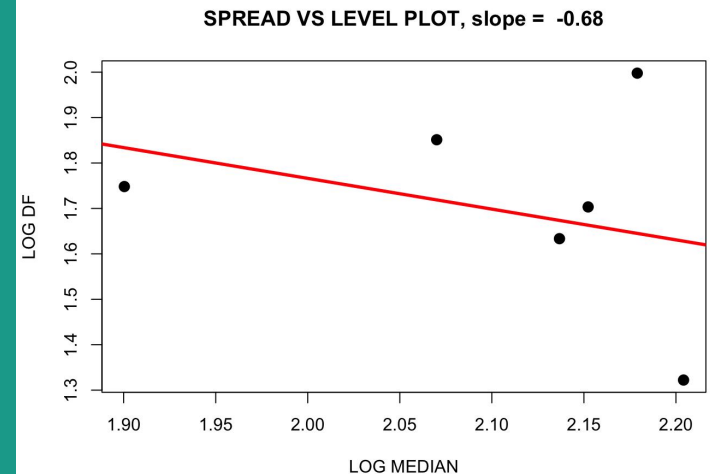


Spread & Level Plot: AtBat & Hits

AtBat: A negative relationship is evident in the graph showing batches with smaller medians tend to have larger spreads (dfs).

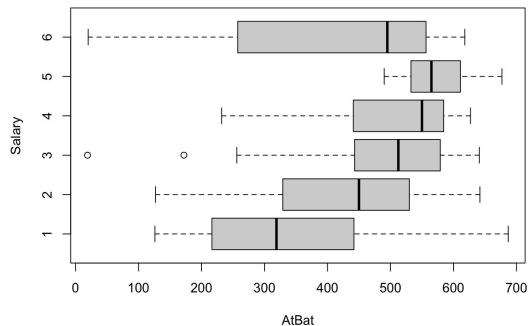
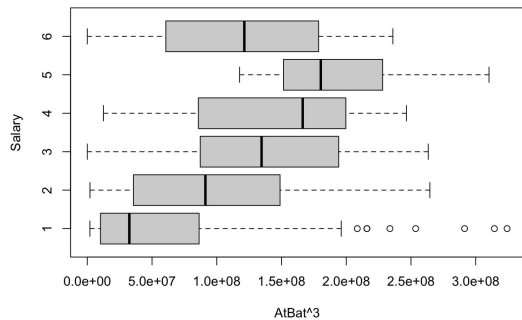


Hits: The dfs (spreads) of batches with small medians are also regarded as small, suggesting a weak negative association. In other words, spreads are dependent on levels.

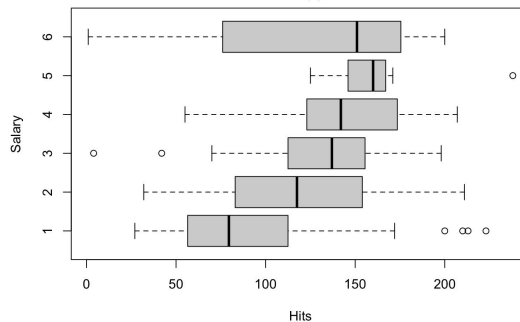
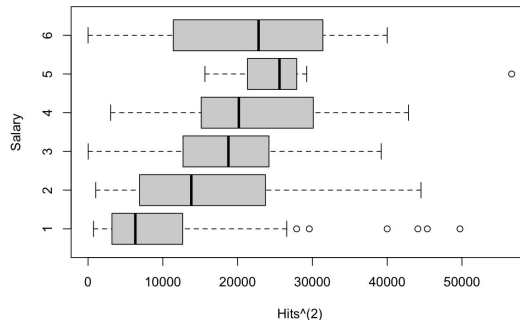


Re-express: AtBat & Hits

AtBat: slope = -1.8;
 $p = 1 - b$; Hence $p = 2.8$; approx. 3



Hits: slope = -0.68;
 $p = 1 - b$; Hence $p = 1.68$; approx. 2



Re-expression for $\text{AtBat}^{(2)}$:

There is a slight improvement in the box plot and as we can see still see that all boxes do not have same spread.

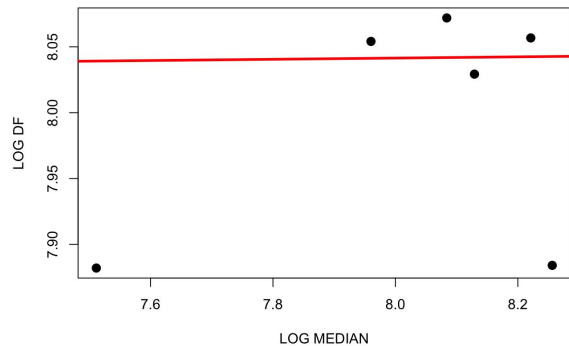
Re-expression for $\text{Hits}^{(-2.5)}$:

The spread of box plots did not change significantly. So, this re-expression does not work well for this batch.

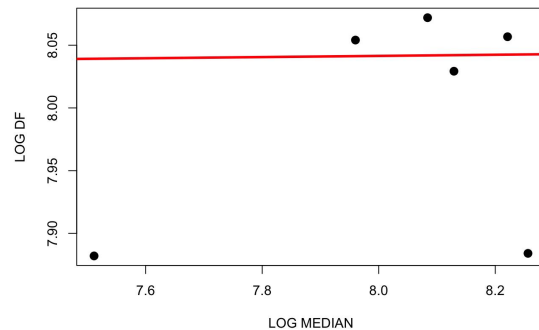
Spread & Level Plots



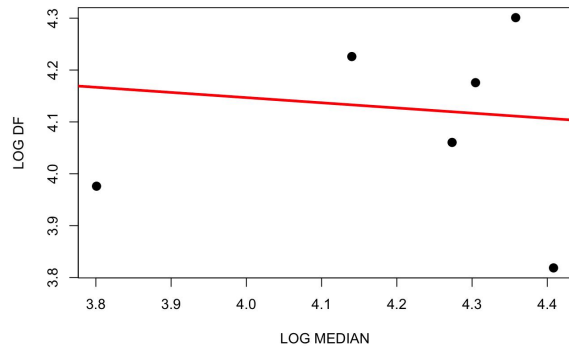
SPREAD VS LEVEL PLOT, slope = 0



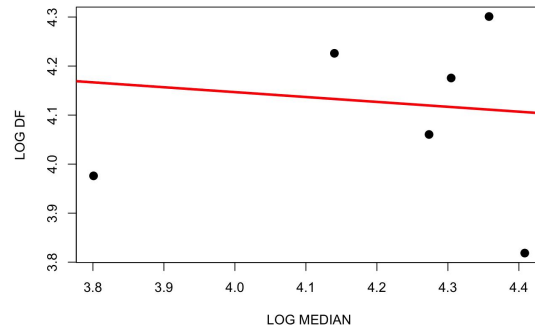
SPREAD VS LEVEL PLOT, slope = 0



SPREAD VS LEVEL PLOT, slope = -0.1

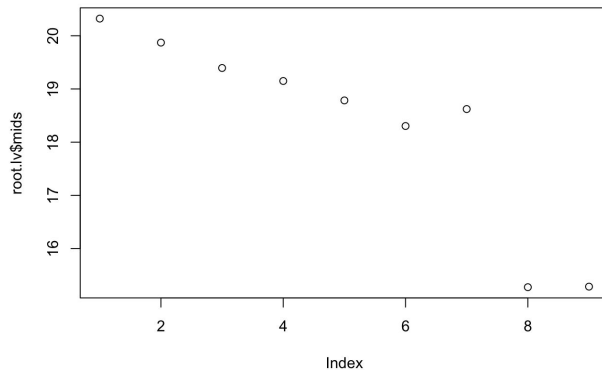


SPREAD VS LEVEL PLOT, slope = -0.1

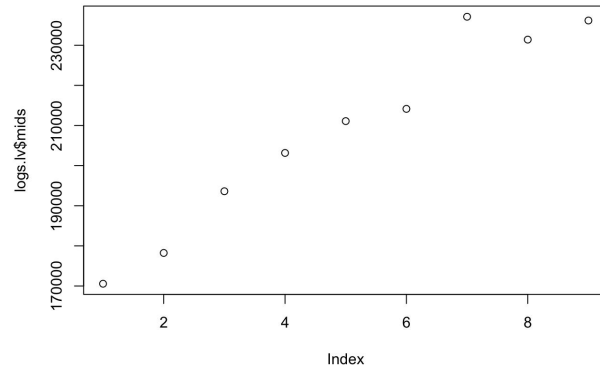


Hinkley's method: AtBat

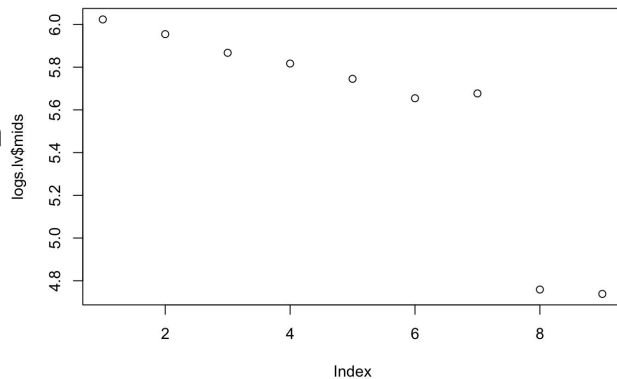
P = 0.5



P = 2



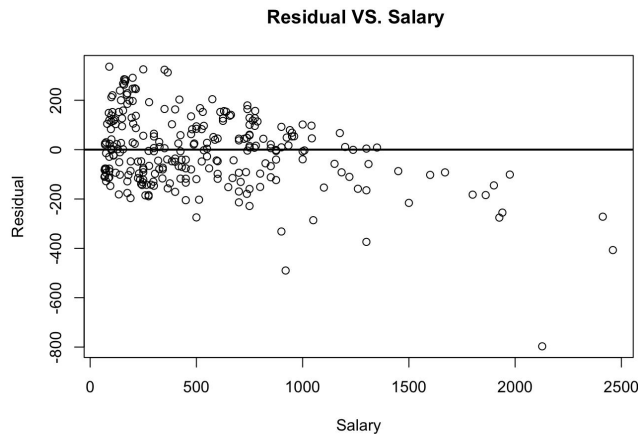
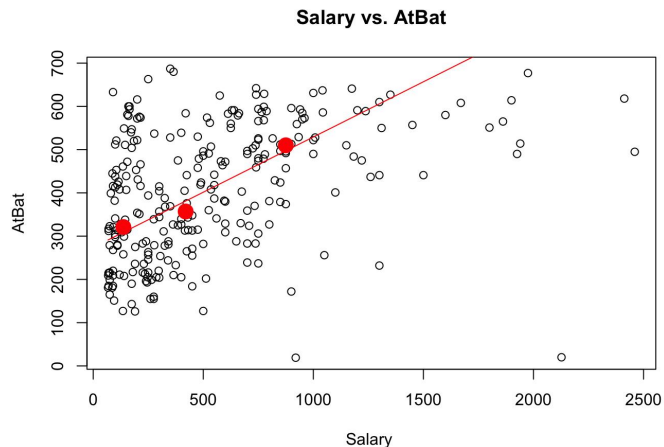
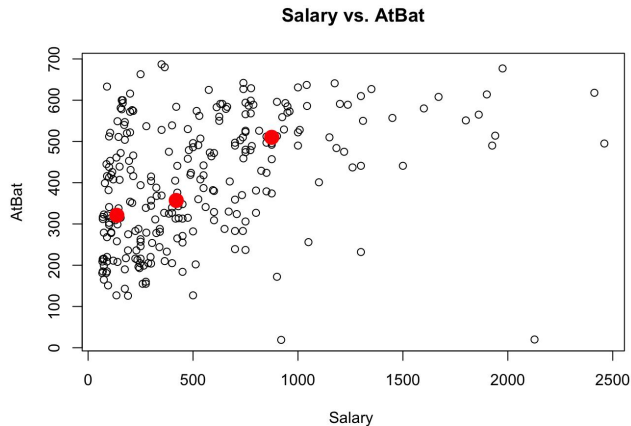
P = 0:
Log Transformation



Resistant fit: Salary vs AtBat



In general, AtBat levels in people have increased at a rate of 0.2552614 per Salary, based on the slope of the line $m = 0.2552614$. There you have it, an explanation of how Salary can affect one's AtBat level. The AtBat increases by 0.2552614 when the Salary is increased by one year.



Conclusion



- The red dots indicate the squaring of the fitted counts (according to the Gaussian distributions). On the fits, we "hang" the actual bin counts (the bars). A deviation from normality occurs if any bars do not land on 0.
- The first thing I notice is that there are many negative deviations, followed by positive deviations. In other words, our data show higher levels of AtBat than we'd expect based on the Gaussian curve.
- On the left side of the plot, there are positive deviations as well. AtBat levels are lower than we would expect if we assume normality.
- The middle of the distribution also exhibits a large negative deviation.
- My conclusion is that the roots of AtBat levels are not quite Gaussian distributed.



Thank You



Scan me!