Analysis Of Hockey Plus-Minus Using Linear Regression

Reading the data from several different csv files. Formatting the data into one data set

```
#load the data and power play data
data <- read.csv("Data/2007-2017.csv")
data_PP <- read.csv("Data/2007-2017_PP.csv")
#create row names
rownames(data) <- (data$Player)
```

```
#match data by player name between files
playerID = match(row.names(data), data_PP$Player)
##add power play data
data$PP_G = data_PP$G[playerID]
data$PP_A = data_PP$A[playerID]
data$PP_G...=data_PP$G...[playerID]
data$PP_GA=data_PP$GA[playerID]
data$PP_GF=data_PP$GF[playerID]
```

```
#Adding salary and Position
##import salary data
salary <- read_csv("Data/Names.csv",
    col_names = FALSE)
playerID = match(row.names(data), salary$X1)
data$Salary = salary$X2[playerID]
##make values numbers
data$Salary = as.numeric(data$Salary)

#seperate Defence from Forwards
Defence <- read_csv("Data/Defence.csv")
##create new column that holds D
Defence$Pos = rep('Defence',length(Defence$Player))
playerID = match(row.names(data), Defence$Player)
data$Pos = Defence$Pos[playerID]
#Identify the forwards
data$Pos[is.na(data$Pos)] <- 'Forward'
new = subset(data, !is.na(data$Salary))
```

Figure 1 is a kernel density plot for players accumulated plus-minus separated by position. The data appears to be relatively normally distributed. Forwards are a bit more skewed.

```
#Density plot for plus-minus by position
caption = "Figure 1: Kernal Density plots of Forward and Defenceman accumulated plus-minus form 2008 to
ggplot(new,aes(G..., colour = Pos, fill = Pos))+
  geom_density(alpha = 0.1)+
  labs(title = "Accumulated Plus-Minus Distribution By Position", caption = "Figure 1: Kernal Density pl
```

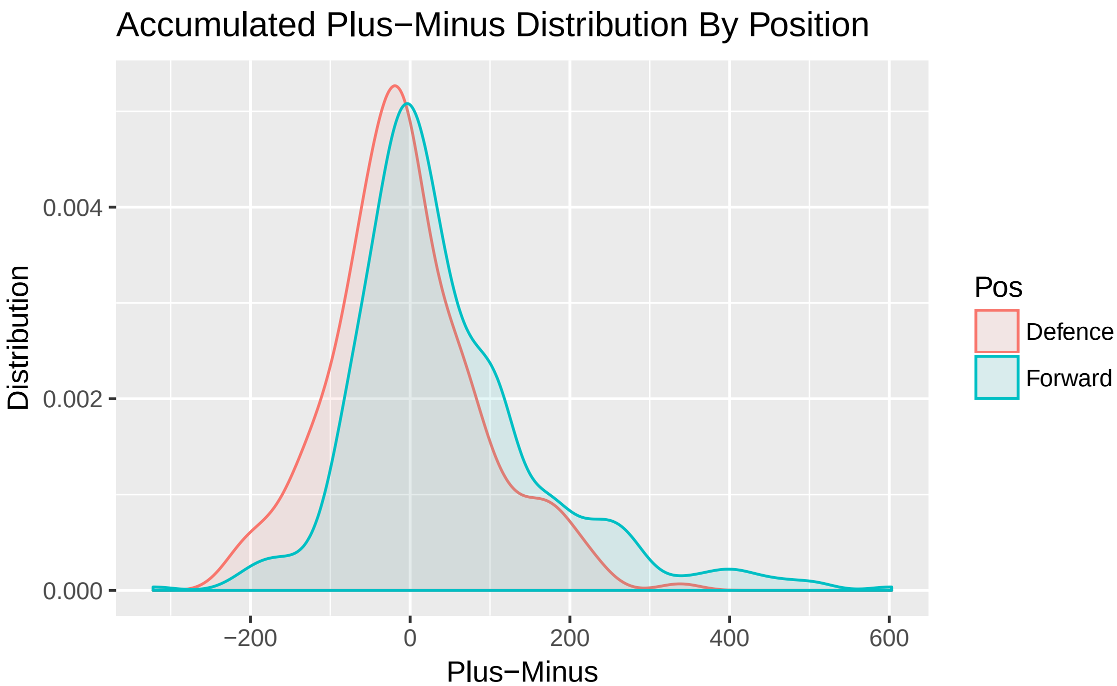

Figure 1: Kernal Density plots of Forward and Defenceman accumulated plus−minus form 2008 to 2017.

Figure 2 is a scatter plot to visualize the relationship between plus-minus and player salary. The data was grouped by position and the salary was set to a log scale. The scatter plot shows a positive association between salary and plus-minus. Model 1 is a linear regression model with response variable as players accumulated plus-minus and response variables salary and position. Model 1 is overlaid on the observed data in figure 1.

Model 1:

$$AccumulatedPlusMinus = \beta_0 + \beta_1 \times Salary + \beta_2 \times Position + \epsilon$$

```
##linearmodel with Salary and Position
lm_Pos = lm(G...~ Pos + Salary,data=new)
```

```
library(stringr)
#Plot accumulated plus-minus agains salary, color by positions and overlay Model 1

#fit the predicted values
pred = data.frame(G... = predict(lm_Pos))
playerID = match(row.names(pred), new$Player)
pred$Salary = new$Salary[playerID]
pred$Pos = new$Pos[playerID]

#plot
caption = "Figure 2: Scatter plot of players accumulated plus-minus form 2008 to 2017 vs accumulated pla

ggplot(data,aes(x=Salary, y=G..., color = Pos))+
  geom_point()+
  ggtitle("Career Plus-Minus vs Salary By Position")+ xlab("Salary") + ylab("PM")+
  scale_x_log10()+
  geom_line(data=pred)+
  ggtitle("Accumulated Plus-Minus vs Salary")+ xlab("Player Salary") + ylab("Accumulated Plus-Minus")+
  labs(caption = str_wrap(caption,120))+theme(plot.caption=element_text(size=9, hjust=0, margin=margin(
```

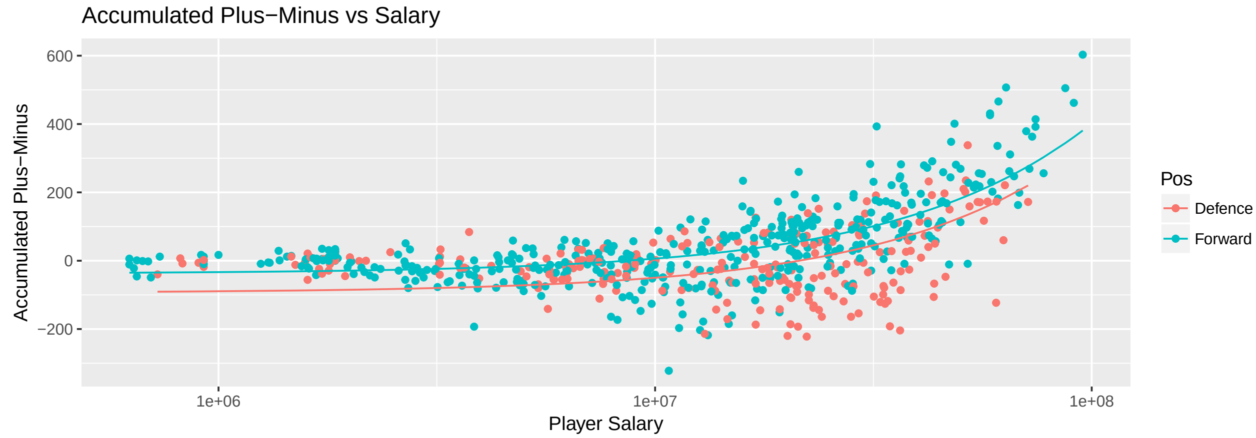

Figure 2: Scatter plot of players accumulated plus−minus form 2008 to 2017 vs accumulated player salary, catagorized by positoin. Model 1 is overlaid on the data, again catagorized by position

Figure 2 shows the residuals from Model 1 plotted against salary. The residuals in Figure 2 do not look healthy, the residuals appear to increase as the salary increases.

```
#summary(lm_Pos)
plot(x=log(new$Salary),y=resid(lm_Pos), main="Residules vs plus-minus",sub="Figure 3: Residules for the
```

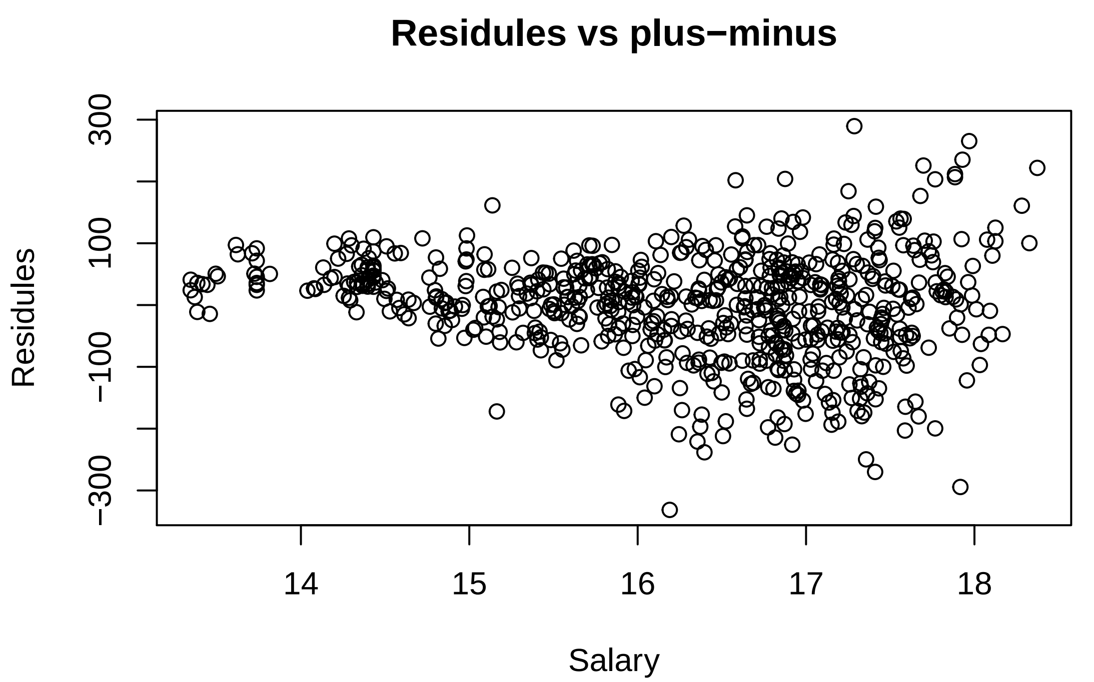

Figure 3: Residules for the linear model against the log scaled salary

```
#plot(density(resid(lm_Pos))) # density distribution, clearly more higher negative residules
#want to look more uniform
#how to deal wwith skew regression coefi
#AIC(lm_Pos) #AIC decreases
```

```
#summary table model 1
lm_Pos %>% tidy(conf.int = TRUE) %>% knitr::kable(digits = 100, col.names = c("Terms", "Estimates", "St
```

Table 1: Table 1: Summary statistics for Model 1

Terms	Estimates	Std Error	Wald Statistic	P-Value	.05 Confidence	.95 Confidence
(Intercept)	-9.379708e+01	6.722286e+00	-13.953152	5.454847e-39	-1.069966e+02	-8.059758e+01
PosForward	5.604918e+01	6.838815e+00	8.195745	1.283956e-15	4.262088e+01	6.947749e+01
Salary	4.390840e-06	1.910186e-07	22.986455	1.807552e-86	4.015767e-06	4.765913e-06

Figure 4 is a kernel density plot for players average plus-minus separated by position. The data appears to be relatively normally distributed. Both distributions could be a mixture of two different distributions.

```
# Average plus-minus per game
#create +/- per game data
data$G...game = data$G.../data$GP
new = subset(data, !is.na(data$Salary))

ggplot(new,aes(G...game, colour = Pos, fill = Pos))+
  geom_density(alpha = 0.1)+
  labs(title = "Average Plus-Minus Distribution By Position", caption = "Figure 4: Kernal Density plots
```

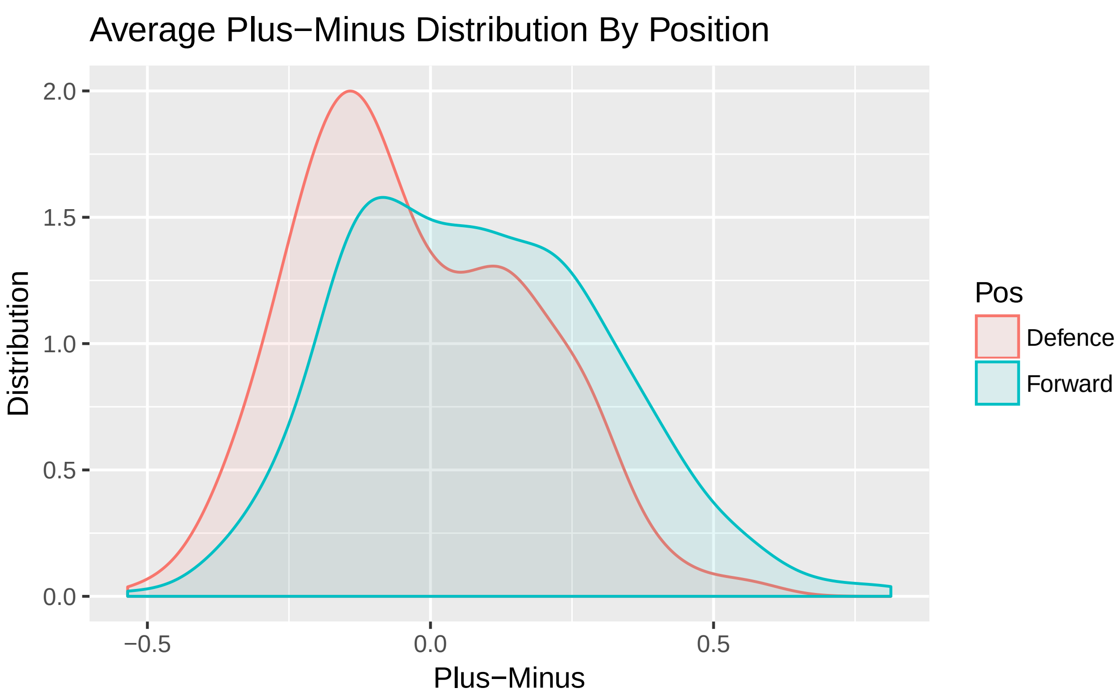

Figure 4: Kernal Density plots of Forward and Defenceman average plus−minus form 2008 to 2017.

Figure 5 is a scatter plot between plus-minus per game and player salary. Again salary is set to a log scale. There is a slight positive relationship between plus-minus per game and salary. Model 2 is a linear regression model with response variable as players plus-minus and response variables salary and position. Model 2 is overlaid on observed data in Figure 5.

Model 2:

$$AveragePlusMinusPerGame = \beta_0 + \beta_1 \times Salary + \beta_2 \times Position + \epsilon$$

```
#Model 2
lm_Pos_game = lm(G...game~Pos+Salary,data=new)
```

```r
#salary vs average plus-minus
playerID = match(row.names(pred), new$Player)
pred$G...game = predict(lm_Pos_game)

#plot
caption = "Figure 5: Scatter plot of players average plus-minus form 2008 to 2017 vs accumulated player

library(stringr)
ggplot(new, aes(x= Salary, y= G...game, color= Pos))+
  geom_point()+
  scale_x_log10()+
  geom_line(data=pred)+
  ggtitle("Average Plus-Minus Per Game vs Salary")+ xlab("Player Salary") + ylab("Average Plus-Minus Pe
  labs(caption = str_wrap(caption,120))+theme(plot.caption=element_text(size=9, hjust=0, margin=margin(
```

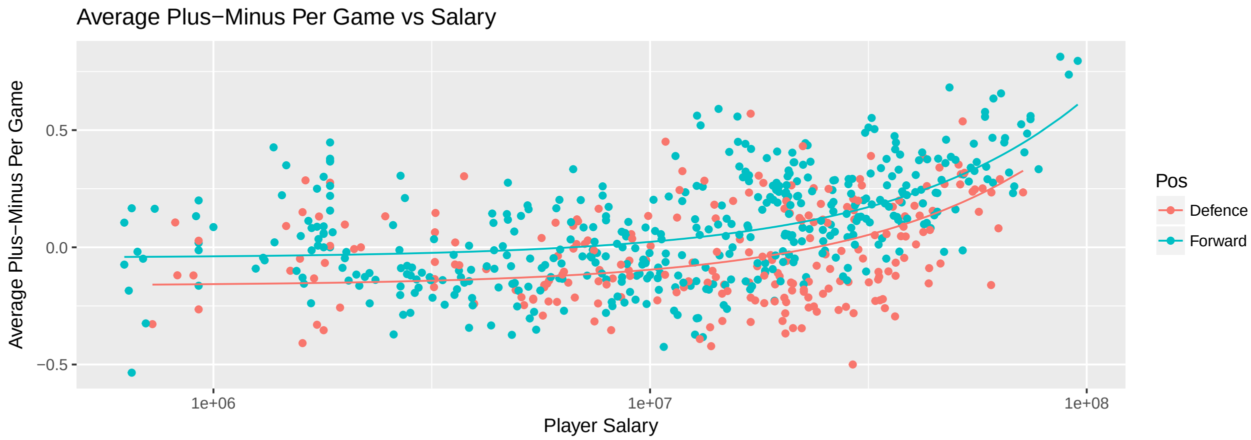

Figure 5: Scatter plot of players average plus−minus form 2008 to 2017 vs accumulated player salary, catagorized by positoin. The linear regression model is overlaid on the data, again catagorized by position

Figure 6 is the Model 2 residuals plotted against player salary. Table 2 summary statistic for Model 2 shows significance for both salary and position coefficients. Both coefficients show a positive relation to average plus-minus.

```r
#model 2 residule plot
plot(x=log(new$Salary),y=resid(lm_Pos_game), main="Residules vs Plus-Minus Per Game", xlab = "Salary", 
```

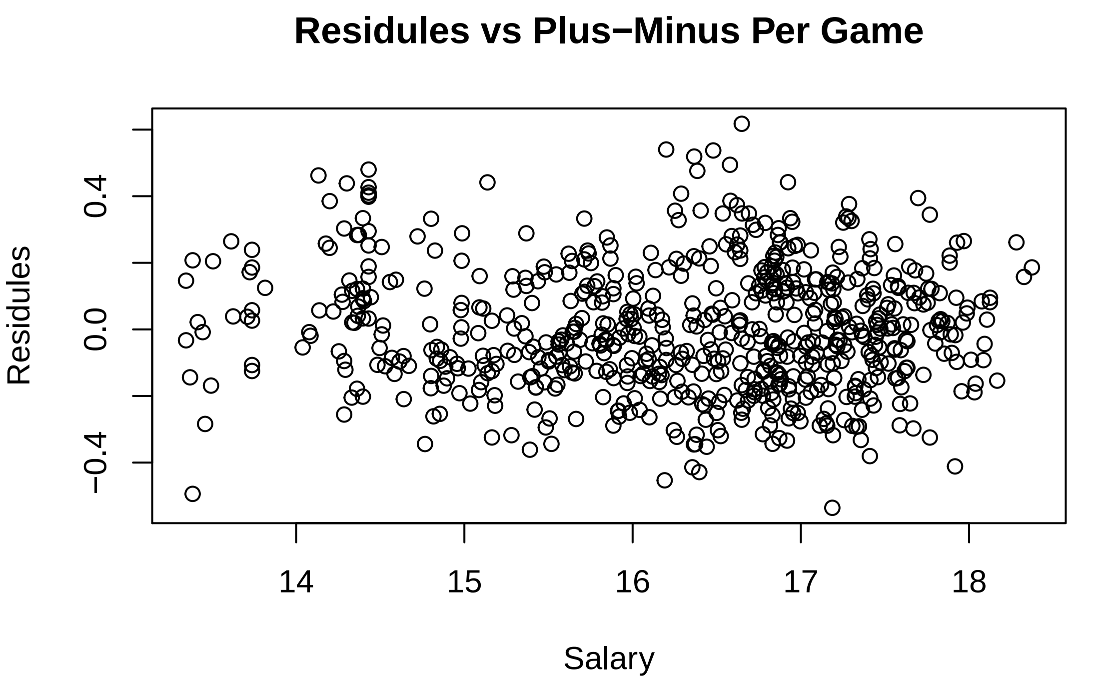

Figure 6: Model 2 residules against the average plus−minus per game

```
#model 2 summary table
tidy(lm_Pos_game, conf.int = TRUE) %>% knitr::kable(digits = 100, col.names = c("Terms", "Estimates", "S
```

Table 2: Summary statistics for linear regression to predict average plus-minus per game

Terms	Estimates	Std Error	Wald Statistic	P-Value	.05 Confidence	.95 Confidence
(Intercept)	-1.640206e-01	1.500709e-02	-10.929542	1.103317e-25	-1.934877e-01	-1.345536e-01
PosForward	1.187330e-01	1.526724e-02	7.776979	2.842930e-14	8.875510e-02	1.487109e-01
Salary	6.862683e-09	4.264373e-10	16.093066	1.905013e-49	6.025355e-09	7.700011e-09

Plus-minus is commonly criticized because of its simplicity and not taking non scoring factors into account. To better understand what variables affect plus-minus, a linear regression model was created with response variable accumulated player plus-minus based on non-scoring related statistics. Figure 7 is a pairwise plot categorized by position (forward in blue, defense in red) containing accumulated player plus-minus and the non scoring related statistics; games played, penalty plus-minus, time on ice percentage, zone start ratio and position. The correlations among the coefficient variables are relatively low, with the exception of salary vs games played. Forward selection and Akaike's Information Index is used select the variables to produce the best fitting linear model. The best fitting model came from using Zone start ratio and games played.

Model 3: $AccumulatedPlusMinus = \beta_0 + \beta_1 \times ZoneStartRatio + \beta_2 \times GamesPlayed + \epsilon$

```
# new variables, non scorign related
pairing = data.frame('Plus_Minus'=new$G...,'Games.Played'=new$GP, 'Penalty.Plus_Minus'= new$iP...,'Time
#pairwise plot
caption = "Figure 7: Pairwise plot of accumulated player plus-minus and non scoring statistical measurme
pairs = ggpairs(pairing,aes(colour=Pos))+
  ggtitle("Parwise Plot Of Non Scoring Variables")+
  labs(caption = str_wrap(caption,120))+theme(plot.caption=element_text(size=9, hjust=0, margin=margin(
print(pairs, progress = FALSE)
```

Figure 7: Pairwise plot of accumulated player plus−minus and non scoring statistical measurments including: Games Played, Peantly Plus−Minus, Percentage Time On Ice, Zone Start Ratio, Salary, and Position. The data is catagorized by player position. The diagnal is the density plots, the lower triangular are the scatterplots, the upper trianglular are the correlations and correlations by position.

Figure 8 is the Model 3 residuals plotted against Zone Start Ratio and Games Played. Both residual plots appear fairly healthy with no definite trends. Table 3 summary statistics for Model 3 shows significance for both Zone start ratio and games played coefficients. Both coefficients show a positive relation to accumulated plus-minus.

```
library(pander)
nonscoringLM = lm(Plus_Minus ~ Zone.Start.Ratio + Games.Played,pairing)
par(mfrow=c(1,2))
plot(x=new$ZSR,y=resid(nonscoringLM), main="Residules vs Plus-Minus Per Game", xlab = "Zone Start Ratio
plot(x=new$GP,y=resid(nonscoringLM), main="Residules vs Plus-Minus Per Game", xlab = "Games Played", yla
```

```
#summary statistics for Model 3
tidy(nonscoringLM, conf.int = TRUE) %>% knitr::kable(digits = 100, col.names = c("Terms", "Estimates",
```

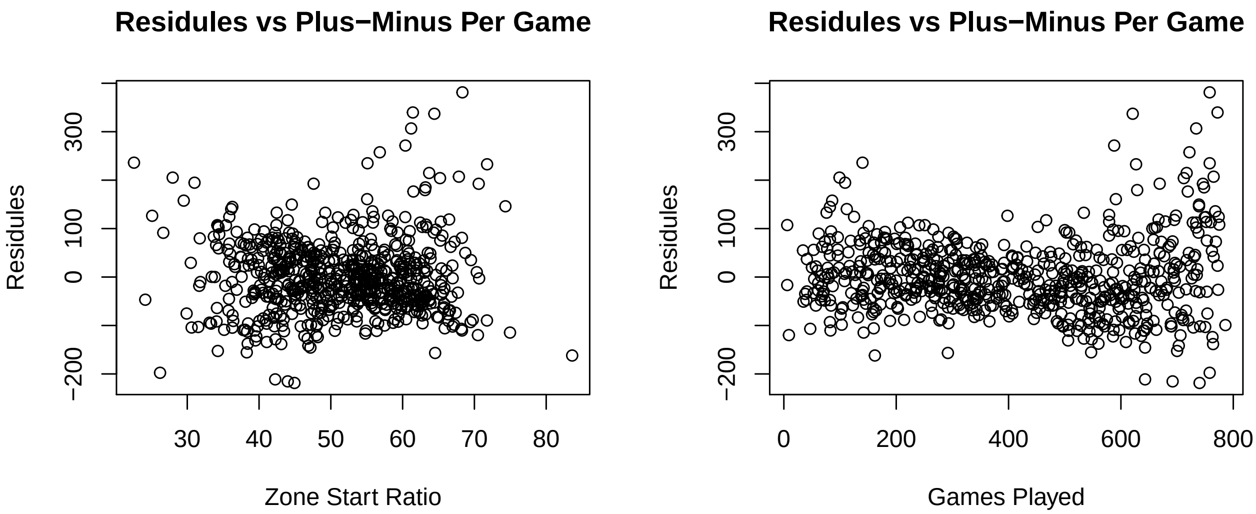

Figure 1: Figure 8: Model 3 residules against Zone Start Ratio and Games Played

Table 3: Table 3: Summary statistics for Model 3 accumulated plus-minus using non scoring stats

Terms	Estimates	Std Error	Wald Statistic	P-Value	.05 Confidence	.95 Confidence
(Intercept)	-459.8023853	16.85230078	-27.28425	0.00000e+00	-492.8926041	-426.7121665
Zone.Start.Ratio	8.2173736	0.30342784	27.08181	0.00000e+00	7.6215800	8.8131673
Games.Played	0.1584141	0.01399287	11.32106	2.72676e-27	0.1309385	0.1858897