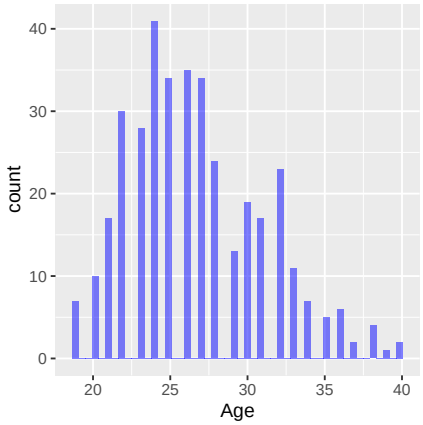
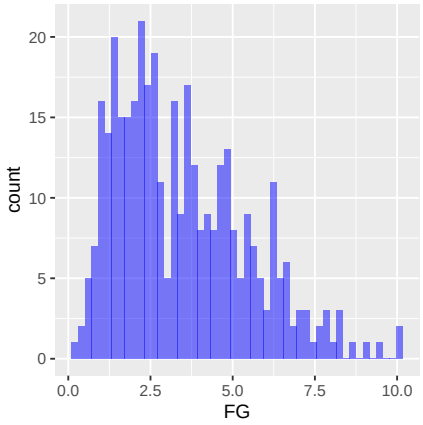
Hot Hands Case Study

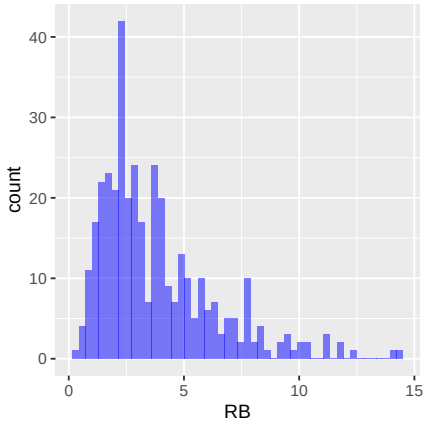
Day Group 4

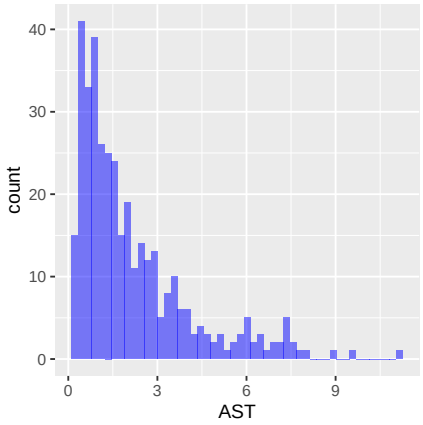
Visualization We will begin by making histograms of `Salary`,`Age`,`FG`,`RB`,`AST`,`STL`,`BLK`, and `PTS`.

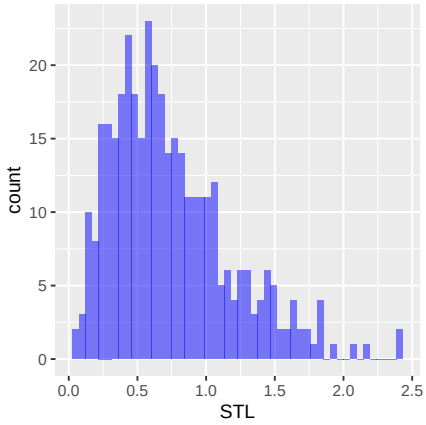

count
Salary
0
10000
20000
0
20
40
60

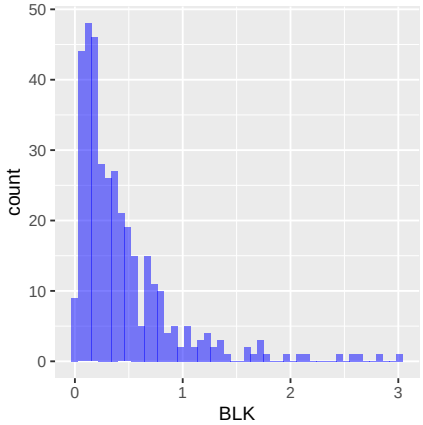

count
Age
20
25
30
35
40
0
10
20
30
40

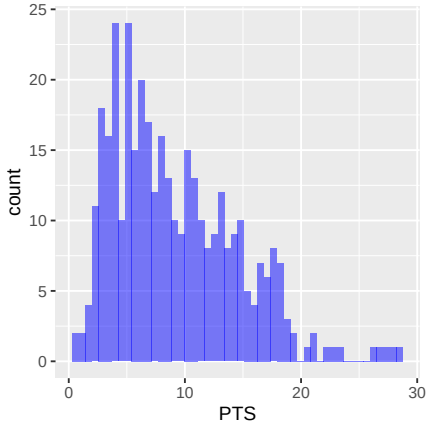

count
FG
0.0
2.5
5.0
7.5
10.0
0
5
10
15
20


count
RB
0
5
10
15
0
10
20
30
40


count
AST
0
3
6
9
0
10
20
30
40

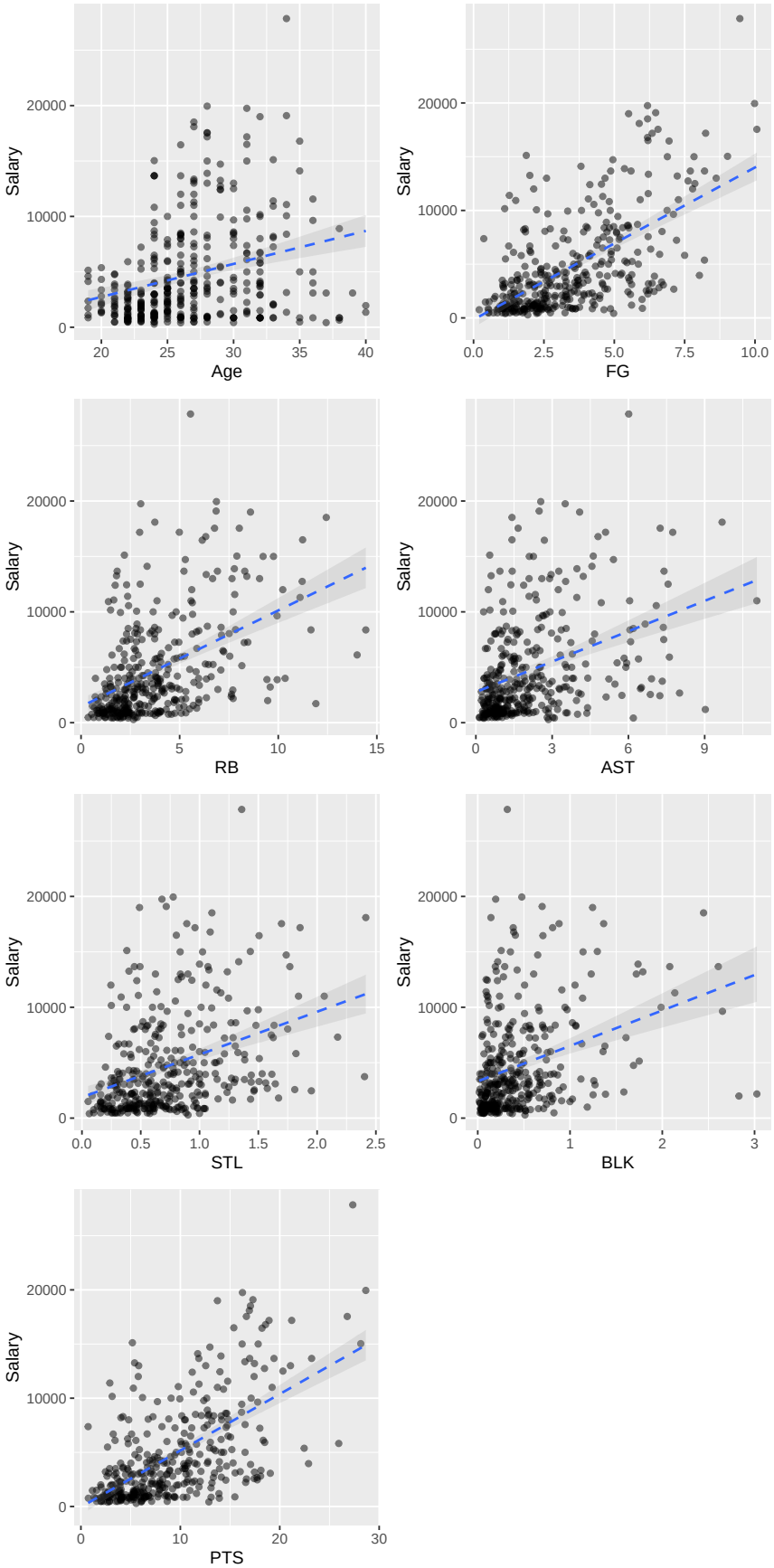

count
STL
0.0
0.5
1.0
1.5
2.0
2.5
0
5
10
15
20


count
BLK
0
1
2
3
0
10
20
30
40
50


count
PTS
0
10
20
30
0
5
10
15
20
25

The histograms above are right skewed.

Scatterplots of `Salary` versus each of the predictors and their relationships are shown below.


Salary
Age
FG
RB
AST
STL
BLK
PTS

The relationship between Salary and all the predictors is positively correlated, with the exception of Age which does not have a visual indicator of correlation since the data is more spread out. These scatterplots also show that the Salary values are in greater volume in the bottom left corner, with outliers present for all the predictors.

Regression A We will run an initial regression of Salary on the predictors:Age,FG,RB,AST,STL, and BLK, leaving out PTS for now.

```
Linear regression (OLS)
Data     : nba_pgdata
Response variable    : Salary
Explanatory variables: Age, FG, RB, AST, STL, BLK
Null hyp.: the effect of x on Salary is zero
Alt. hyp.: the effect of x on Salary is not zero

             coefficient std.error t.value p.value
 (Intercept)   -8724.667  1131.833  -7.708  < .001 ***
 Age             312.092    39.520   7.897  < .001 ***
 FG             1156.982   153.881   7.519  < .001 ***
 RB              223.312   117.004   1.909   0.057 .
 AST             280.498   146.504   1.915   0.056 .
 STL           -1064.070   613.248  -1.735   0.084 .
 BLK            1071.100   517.030   2.072   0.039 *

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.492,  Adjusted R-squared: 0.484
F-statistic: 58.632 df(6,363), p.value < .001
Nr obs: 370
```

Interpretation of the estimated coefficients:
The H_0: The effect of the predictor on Salary is 0 (ie. there is no effect)
The H_a: The effect of the predictor on Salary is not 0 (ie. there exists an effect)

1) The predictors that are significant at the 95% confidence level are Age, FG, and BLK since the p-values for these 3 predictors are less than 0.05. Age and FG have a much more significant relationship with Salary than BLK since the p-value is much smaller (< 0.001) for these 2 predictors.

2) These regression results make sense when we consider each of the predictors.

Age: As a younger player on the team, you are likely to earn less than a player that is older, and viewed as an asset for the team. But there is a threshold, at which point there would be a drop-off in Salary since the oldest players are not assets for the team and would likely be paid less.

FG: The correlation between field goals and Salary would be postive since a greater number of goals would likely result in a higher Salary.

BLK: The correlation between blocks and Salary exists as well since the number of blocks would be an indicator for how valuable the player is on the team, thus affecting Salary.

BLK is less common than a rebound (3.78) or assists (2.148), steal (0.739) and block (0.442) as seen by the means (in parantheses). Therefore, we would expect to see a greater significance between Salary and the former 3 predictors, but we observe a significance with the latter.

3) The null and alternative hypotheses above lend themselves to a Chi-squared test.

4) Each unit of increase in the predictors would lead to a Salary change equivalent to the coeffecient corresponding to the predictor.

5) `FG` is the most impactful predictor since it has the highest absolute coefficient value, thus leading to the greatest `Salary` change, followed by `BLK`.

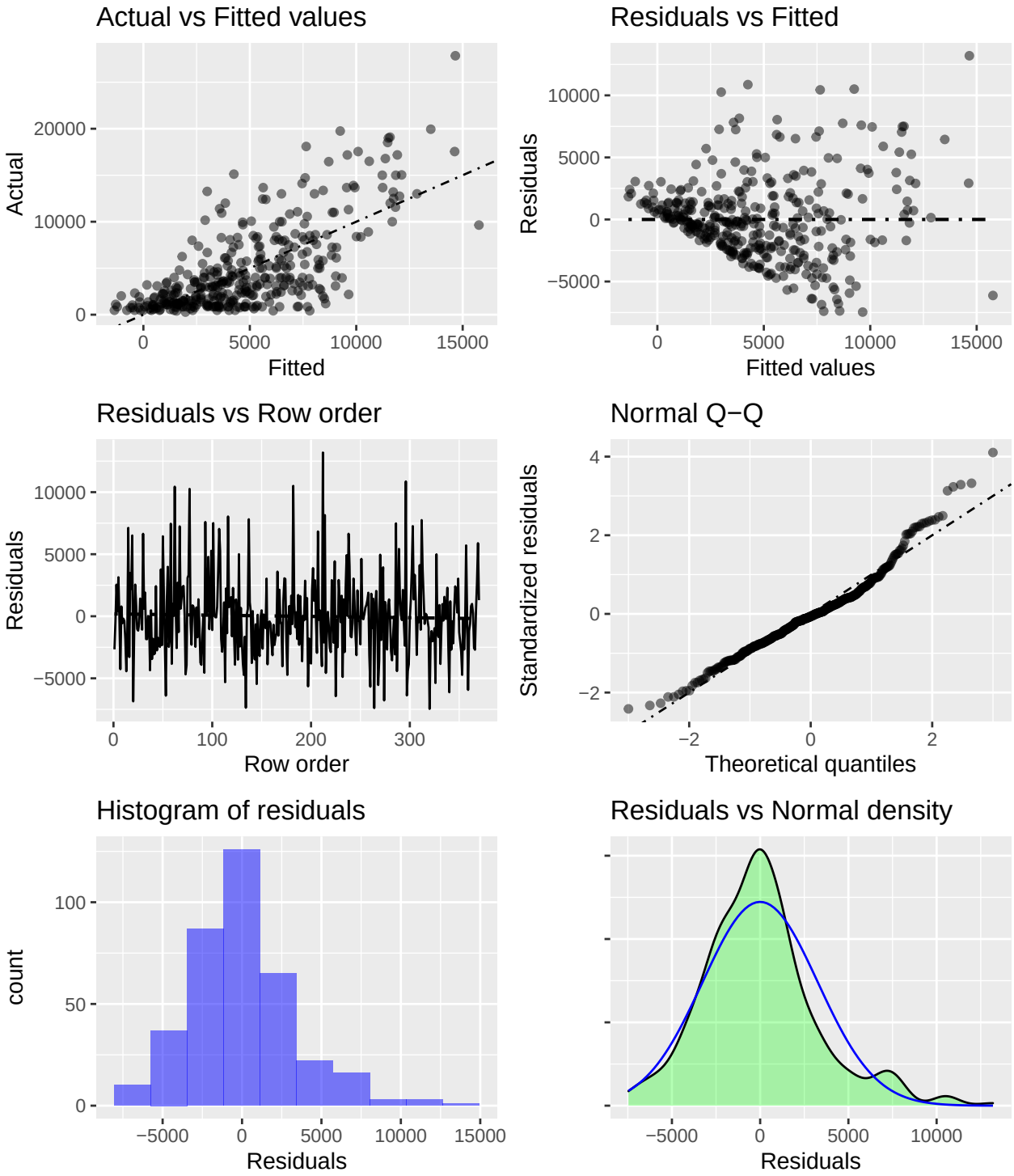
6) The R^2 value is `0.492`, which allows us to interpret how likely this entire set of predictors will affect `Salary`. With the R^2 value of `0.492`, this model explains a 49.2% variation between `Salary` and our regression model.

Since the R^2 value is `0.492`, and the adjusted R^2 value is less, `0.484`, we can assume that there is noise in the predictors that are being tested with this regression model.

```
Linear regression (OLS)
Data     : nba_pgdata
Response variable    : Salary
Explanatory variables: Age, FG, BLK
Null hyp.: the effect of x on Salary is zero
Alt. hyp.: the effect of x on Salary is not zero

             coefficient std.error t.value p.value
 (Intercept)   -8980.906  1119.463  -8.023  < .001 ***
 Age             320.605    39.455   8.126  < .001 ***
 FG             1308.328    92.754  14.105  < .001 ***
 BLK            1449.757   395.495   3.666  < .001 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.483,  Adjusted R-squared: 0.479
F-statistic: 113.925 df(3,366), p.value < .001
Nr obs: 370
```

If the predictors that are statistically insignificant are removed from the model, we can observe the R^2 value does not drastically change, thus leading us to believe that the 3 predictors (`RB`, `AST`, and `STL`) are only adding noise to our regression model.

We will then take a look at the `regression plots`.

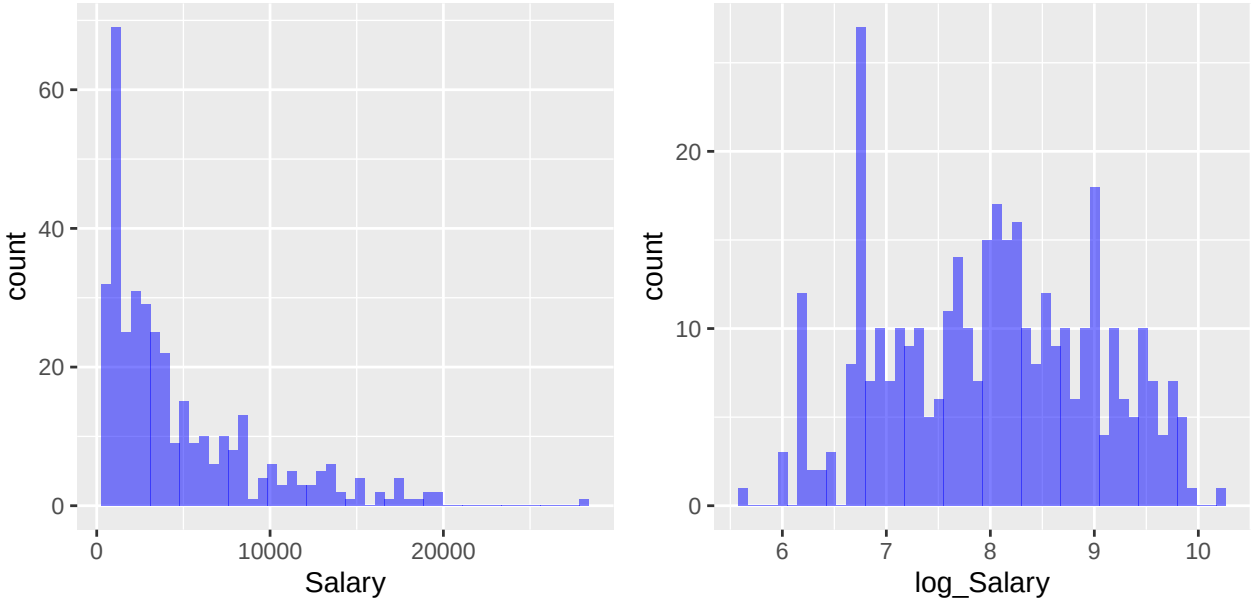


A well behaved residual plot should have a `Residuals vs. Fitted` line that oscillates around the line $y = 0$. But we observe a negative correlation between the 2 values. We also observe that the `Actual vs. Fitted` plot does not closely follow a linear line of fit.

Thus, both of these plots lead us to believe that the relationship between `Salary` and the predictors may not be linear since we are over-estimating some data points and under-estimating others based on our regression model.

Regression B A new variable `log_Salary` is the logarithmic form of the existing `Salary` values.

We will start by visualizing a histogram of `log_Salary` and `Salary`.



From the histograms above, the `log_Salary` histogram solves the problem of skew from the old `Salary` histogram, and with the exception of a few outliers, follows a normal distribution.

We will then run an initial regression of `log_Salary` on the predictors:`Age`,`FG`,`RB`,`AST`,`STL`, and `BLK`, leaving out `PTS` for now.

```
Linear regression (OLS)
Data     : nba_pgdata
Response variable    : log_Salary
Explanatory variables: Age, FG, RB, AST, STL, BLK
Null hyp.: the effect of x on log_Salary is zero
Alt. hyp.: the effect of x on log_Salary is not zero

             coefficient std.error t.value p.value
 (Intercept)       5.302     0.255  20.803  < .001 ***
 Age               0.058     0.009   6.487  < .001 ***
 FG                0.208     0.035   5.994  < .001 ***
 RB                0.067     0.026   2.533   0.012 *
 AST               0.046     0.033   1.403   0.162
 STL               0.002     0.138   0.013   0.990
 BLK               0.210     0.116   1.807   0.072 .

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.455,  Adjusted R-squared: 0.446
F-statistic: 50.599 df(6,363), p.value < .001
Nr obs: 370
```

Interpretation of the estimated coefficients:
The H_0: The effect of the predictor on `log_Salary` is 0 (ie. there is no effect)
The H_a: The effect of the predictor on `log_Salary` is not 0 (ie. there exists an effect)

1) The significant predictors are still `Age`, `FG`, but `RB` is now significant as well. `BLK` is no longer significant at the 95% confidence level since the p-value > 0.05. `RB` is more significant than `BLK` was before. `AST`

and STL are still insignificant at the 95% confidence level since their p-values are greater than 0.05.

2) According to the means from **Regression A**, these results make slightly more sense since the occurrence of a rebound is more than a block, thus having a more significant effect on the Salary. The other 2 metrics, Age and FG being significant are in accordance to our previous reasoning.

3) The null and alternative hypotheses above lend themselves to a Chi-squared test.

4) The estimated coefficients specify the percentage change in Salary per unit change of the predictor (ie. a 1 unit change in Age would result in a 5.8% change in log_Salary).

5) Although BLK has the highest coefficient, the p-value for that predictor was greater than 0.05, thus making it statistically insignificant at the 95% confidence level. Thus, the predictor that is most impactful, and significicant, is still FG, since it has the highest coeffient, followed now by RB.

6) The R^2 value is 0.455, which allows us to interpret how likely this entire set of predictors will affect log_Salary. With the R^2 value of 0.455, this model explains a 45.5% variation between log_Salary and our regression model.

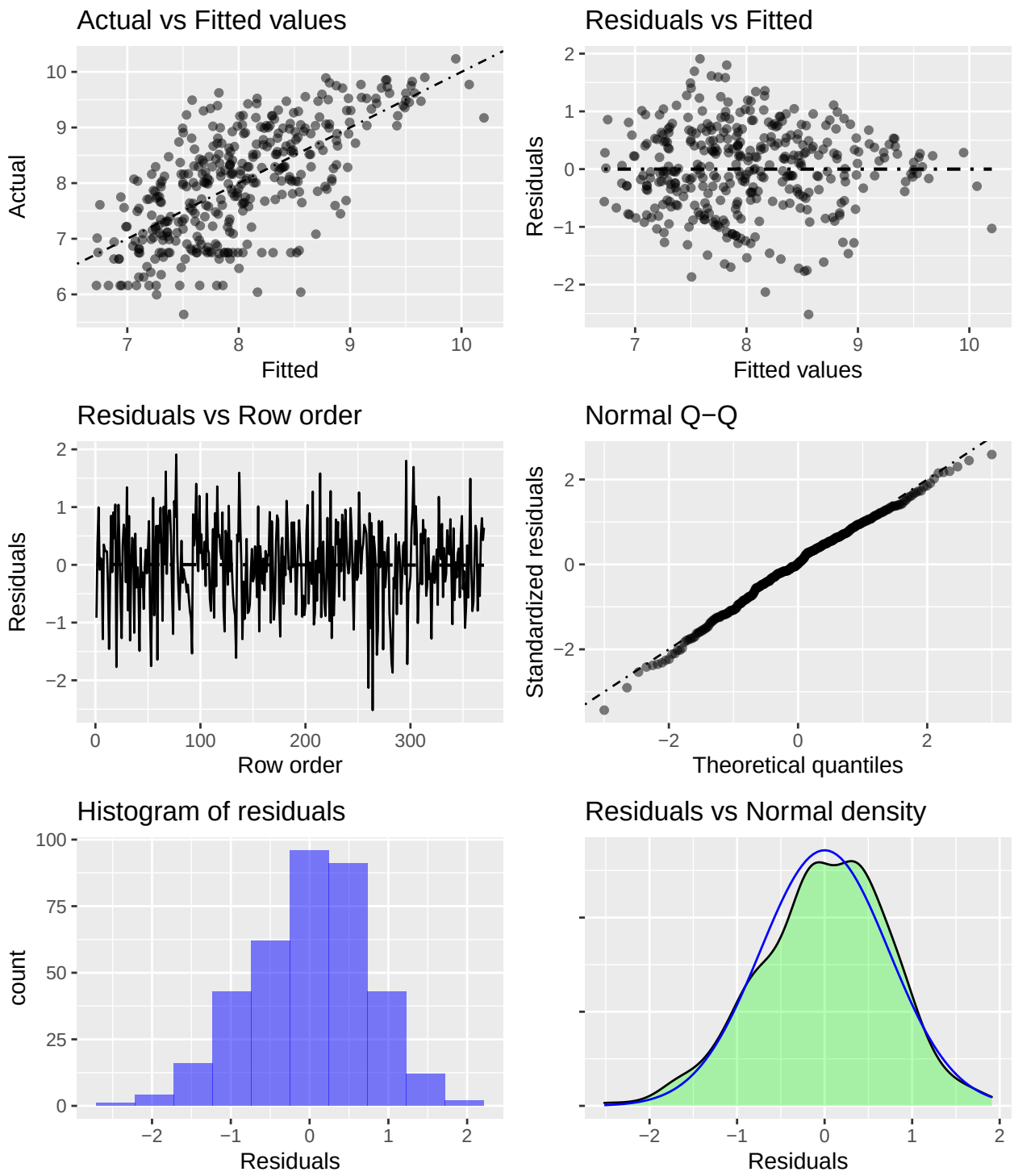
We can see that the data fits this regression model less precisely than the first model, which had an R^2 value of 0.492. Since this R^2 value for this regression is less than **Regression A**, we would likely see a greater difference between observed data and the fitted values. In addition, a lower R^2 value for this regression would indicate that this regression model has a poorer fit to our observed data.

The adjusted R^2 value for this regression is less, 0.446, thus, we can assume that there is noise in the predictors that are being tested with this regression model as well.

```
Linear regression (OLS)
Data     : nba_pgdata
Response variable    : log_Salary
Explanatory variables: Age, FG, BLK
Null hyp.: the effect of x on log_Salary is zero
Alt. hyp.: the effect of x on log_Salary is not zero

             coefficient std.error t.value p.value
 (Intercept)       5.324     0.252  21.089  < .001 ***
 Age               0.059     0.009   6.660  < .001 ***
 FG                0.276     0.021  13.175  < .001 ***
 BLK               0.337     0.089   3.773  < .001 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.444,  Adjusted R-squared: 0.439
F-statistic: 97.372 df(3,366), p.value < .001
Nr obs: 370
```

If the predictors that are statistically insignificant are removed from the model, we can observe the R^2 value does not drastically change, and is closer to the adjusted R^2 value from above, thus leading us to believe that the 3 predictors (RB, AST, and STL) are only adding noise to our regression model. In addition, when these 3 predictors are removed, the p-values for the remaining predictors are all < 0.001, denoting more statistical significance.

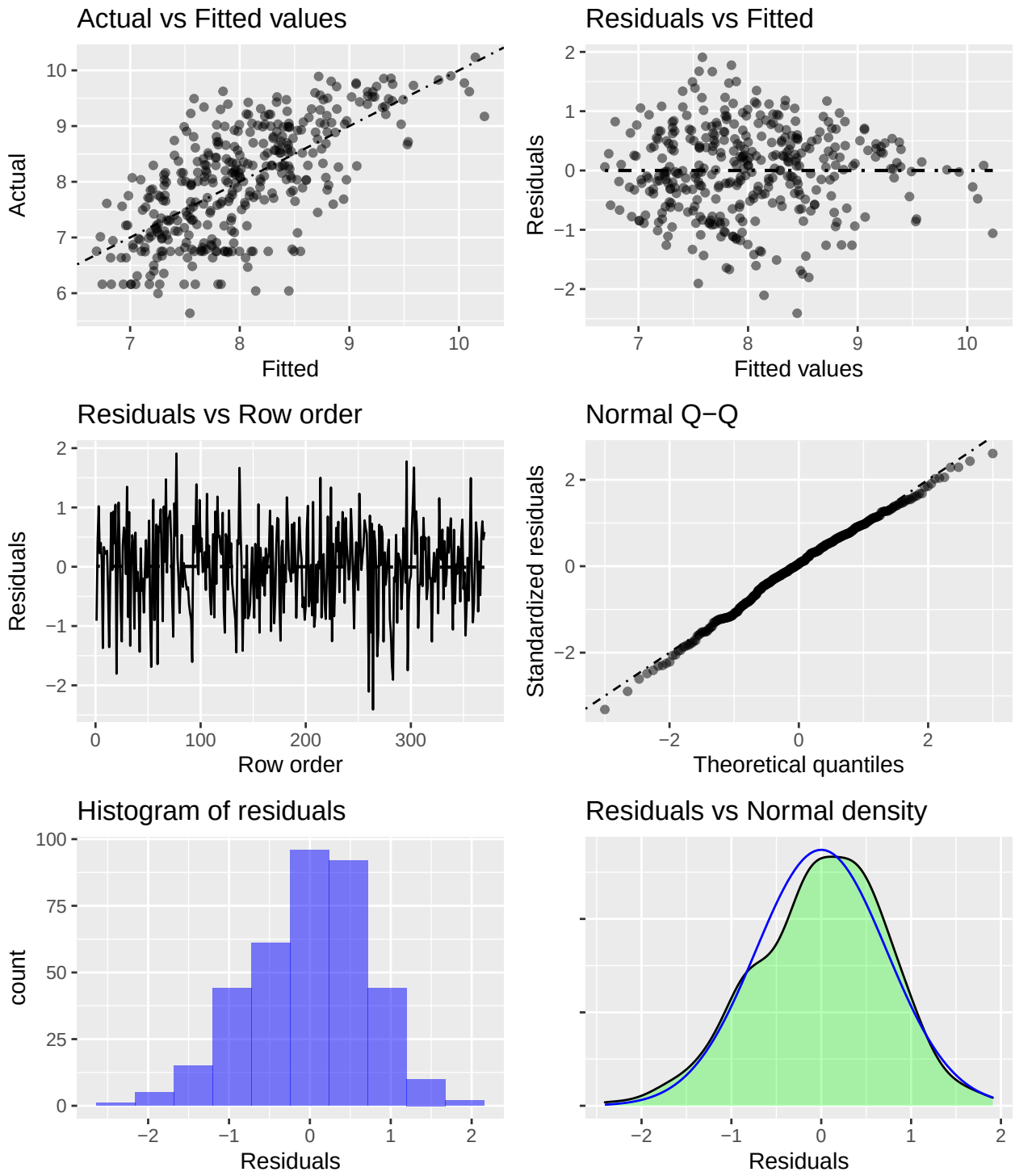
We will now take a look at the regression plots.



From the plots above, the `Actual vs. Fitted` values follow a linear correlation. And it can also be observed that the `Residuals vs. Fitted` plot has values that are more evenly distributed around the $y = 0$ line now, indicating that this is a 'well-behaved' residual plot. This supports our assumption that the relationship is linear between `log_Salary` and the predictors. But since there are values that fall way below -1 and way above 1, we can be certain that there are still outliers in our regression model.

Regression C Keeping the `log_Salary`, we will now add the `PTS` predictor and run our regression model.

```
Linear regression (OLS)
Data     : nba_pgdata
Response variable    : log_Salary
Explanatory variables: Age, FG, RB, AST, STL, BLK, PTS
Null hyp.: the effect of x on log_Salary is zero
Alt. hyp.: the effect of x on log_Salary is not zero

             coefficient std.error t.value p.value
 (Intercept)       5.297     0.253  20.944  < .001 ***
 Age               0.058     0.009   6.524  < .001 ***
 FG               -0.117     0.130  -0.901   0.368
 RB                0.083     0.027   3.081   0.002 **
 AST               0.044     0.033   1.347   0.179
 STL              -0.052     0.139  -0.378   0.706
 BLK               0.251     0.117   2.153   0.032 *
 PTS               0.120     0.046   2.592   0.010 **

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.465,  Adjusted R-squared: 0.455
F-statistic: 45.013 df(7,362), p.value < .001
Nr obs: 370
```



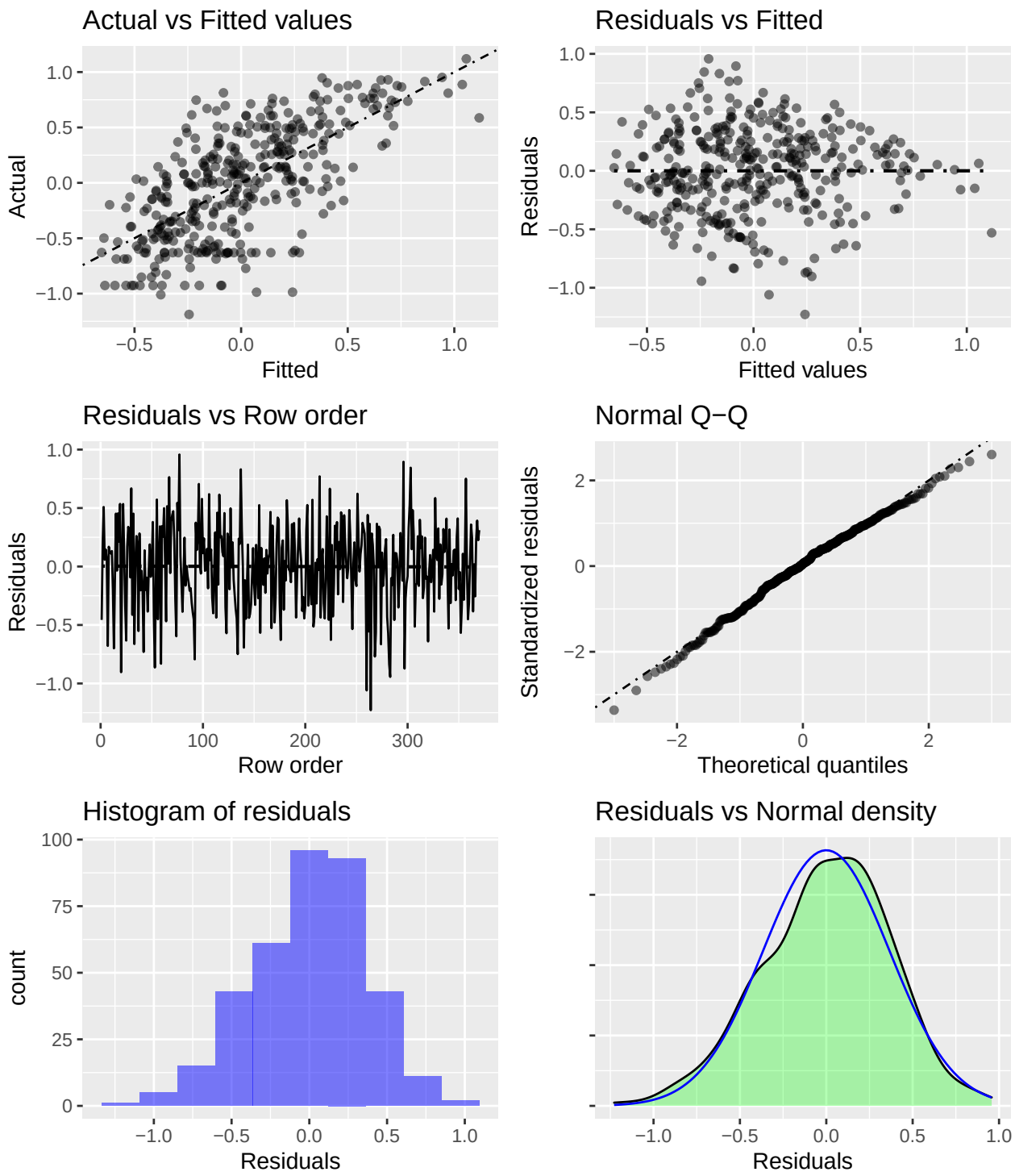
The coefficient for `FG` decreases to a negative value since the additional `PTS` predictor accounts for points from `free-throws` and `field goals`. To ensure there is no double counting, the `FG` coefficient then becomes negative. The addition of `PTS` also makes the `FG` predictor statistically insignificant, but makes the `BLK` predictor significant again and the newly added `PTS` predictor is also statistically significant. This can be explained by the fact that `PTS` refers to points, as opposed to factors of `goals` made. Thus, this predictor would be more directly comparable to `log_Salary`.

The residual plots for this regression do not greatly differ from the previous plots with `FG` alone and no `PTS` predictor. The `Actual vs. Fitted` plot still suggests linearity and the `Residuals vs. Fitted` plot, with

values oscillating around the $y = 0$ line indicate a linear regression model.

Regression D Keeping the `log_Salary` and `PTS` as a predictor, we will now remove `FG` as a predictor. Furthermore, we will standardize the coefficients in the regression model.

```
Linear regression (OLS)
Data     : nba_pgdata
Response variable    : log_Salary
Explanatory variables: Age, RB, AST, STL, BLK, PTS
Null hyp.: the effect of x on log_Salary is zero
Alt. hyp.: the effect of x on log_Salary is not zero
**Standardized coefficients shown (2 X SD)**

             coefficient std.error t.value p.value
 (Intercept)      -0.000     0.019  -0.000   1.000
 Age               0.254     0.039   6.544  < .001 ***
 RB                0.185     0.063   2.953   0.003 **
 AST               0.079     0.062   1.266   0.206
 STL              -0.018     0.060  -0.301   0.763
 BLK               0.111     0.054   2.045   0.042 *
 PTS               0.428     0.066   6.513  < .001 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.464,  Adjusted R-squared: 0.455
F-statistic: 52.407 df(6,363), p.value < .001
Nr obs: 370
```



This change makes both `Age` and `PTS` the most statistically significant predictors since their p-values are both less than `0.001`. `RB` and `BLK` are also significant predictors, but `AST` and `STL` are still statistically insignifiant according to this regression model.

The standardized coefficient indicate that for every change in predictors that is equivalent to 2 standard deviations, the `Salary` will change by the percentage indicated (ie. if `Age` increases by 2 standard deviations, the `Salary` will increase by `25.4%`). Using standardized coeffecients makes the comparisons more direct since the independent variables are all being compared on an identical unitless scale. For example, the comparison now, with standardized coefficients, allows a comparsion to be made between `Age` and `BLK` even though these

two categories are on different scales, since we have reduced them down to simple standard deviations.

The residual plots for this regression do not differ too much from the previous plot with `FG` as a predictor. The linearity of the regression model holds true here as well.

Hot Hand Hypothesis Testing

Conditional Probability Approach Using Table 1 provided in the write up, the following means can be extracted.

```
Explore
Data        : Players
Functions   : n_obs, mean, min, max, sd
Top         : Function

 variable n_obs  mean   min   max    sd
    P_H_M     9 0.547 0.460 0.710 0.077
      P_H     9 0.517 0.460 0.620 0.052
    P_H_H     9 0.502 0.430 0.570 0.048
```

From a bird's eye view: The average probability of a hit given a previous miss (`P_H_M`) is higher than the other two average probabilities.

The idea of a Hot Hand is that you hit more if you have previously hit, rather than if you have previously missed. Therefore, we know that for this phenomenon to exist, the probability of a hit given a previous hit (`P_H_H`) must be greater than the probability of a hit (`P_H`), or the probability of a hit given a previous miss (`P_H_M`).

At first glance, from the table above, we know that these conditions are not satisfied with our current data set, because `P_H_M` $>$ `P_H_H` and `P_H` $>$ `P_H_H`, thus likely disproving the existence of the Hot Hand phenomenon.

Just from examining the data at hand, the only player that seemingly exhibits Hot Hand is Player B, whose probability of hitting after a previous hit (`P_H_H`) is greater than both the probability of a hit (`P_H`) and the probability of a hit given a miss (`P_H_M`).

We can illustrate this statistically as well, with the following two null hypotheses.

H_0: P{H|H} = P{H|M}
H_a: P{H|H} $>$ P{H|M}

H_0: P{H|H} = P{H}
H_a: P{H|H} $>$ P{H}

The sample size is unknown, but the data set contains 9 players from 1 NBA team. The scope of the statistics is also unknown since we do not know whether the probabilities are for 1 season or over multiple seasons.

The most appropriate distribution would be a t-distribution since there are only 9 samples in this data set. Looking at the distribution for this data, it can also be observed that there is a sudden peak, and then the curve flattens, lending itself to a t-distribution.

Using this limited data, a Pairwise mean comparisons test can be conducted, which results in p-values of `0.725` and `0.917`. At the 95% confidence level, the null hypotheses cannot be rejected since `0.725 > 0.05` and `0.917 > 0.05`.

With such a small sample size, a confident conclusion cannot be made regarding the existence of the Hot Hand, although with the limited data available, the null hypothesis cannot be rejected.

```
Pairwise mean comparisons (t-test)
Data      : Players
Variables : P_H_H, P_H_M, P_H
Samples   : independent
```

```
Confidence: 0.95
Adjustment: None

        mean n n_missing    sd    se    me
 P_H_H 0.502 9         0 0.048 0.016 0.037
 P_H_M 0.547 9         0 0.077 0.026 0.059
   P_H 0.517 9         0 0.052 0.017 0.040

 Null hyp.       Alt. hyp.       diff   p.value
 P_H_H = P_H_M   P_H_H > P_H_M   -0.044 0.917
 P_H_H = P_H     P_H_H > P_H     -0.014 0.725

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Streaks Approach The following simulation model was developed to find the `expected number of runs` and `standard deviation of runs`.

```
# A tibble: 9 x 3
  Player Actual_Runs Expected_Runs
  <chr>        <dbl>         <dbl>
1 A              128           125
2 B              431           442
3 C              203           209
4 D              172           168
5 E              134           137
6 F              245           225
7 G              227           216
8 H              176           176
9 I              220           191
```

From the output table above (in which the left column refers to the `expected number of runs` and the right column refers to the `standard deviation`), we can create a hypothesis test.

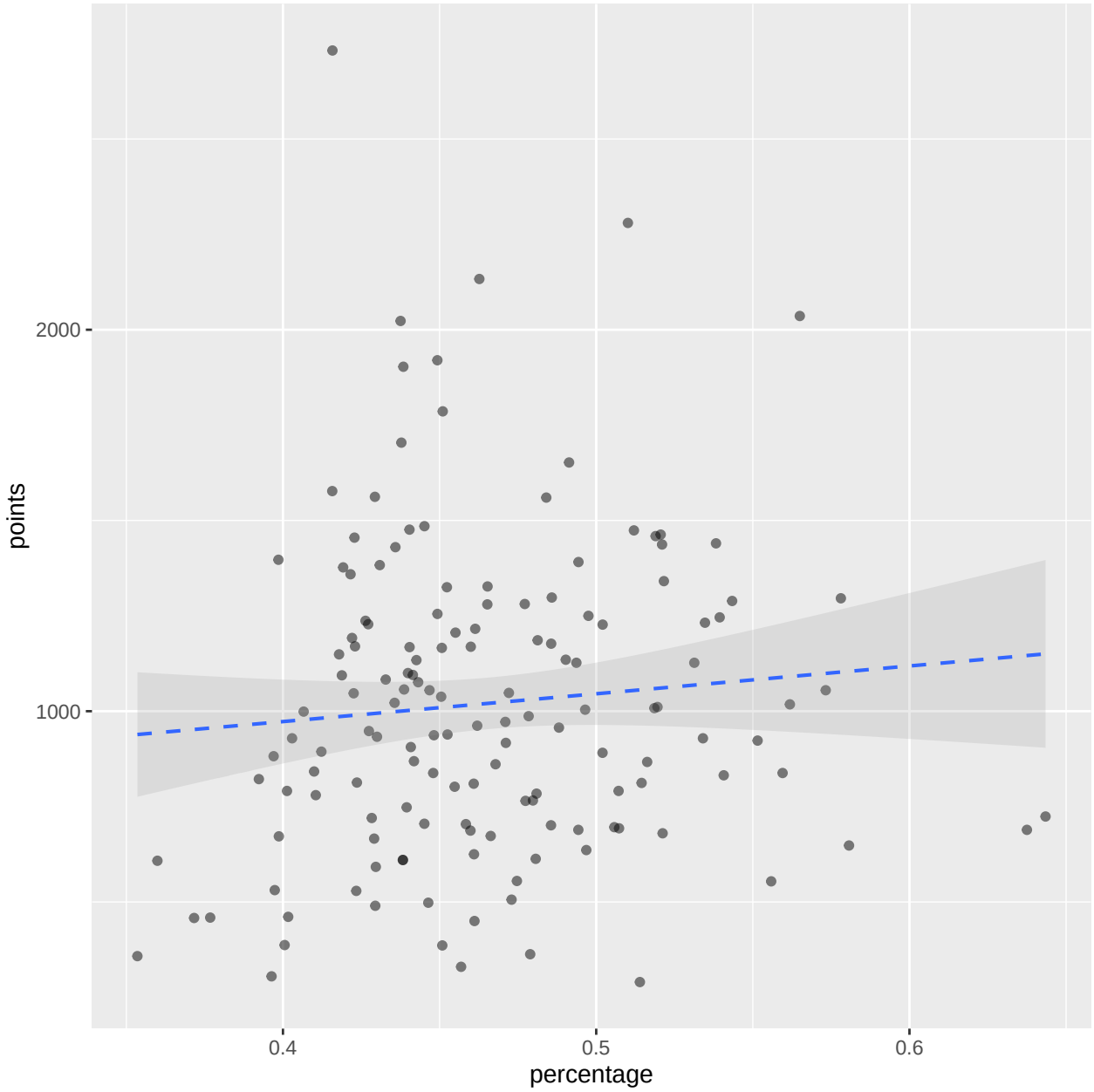
H_0: Actual Runs = Expected Runs
H_a: Actual Runs < Expected Runs

The Actual Runs is less than the Expected Runs for 3 players (`B`, `C` and `E`).
The Actual Runs is greater than the Expected Runs for 5 players (`A`, `D`, `F`, `G` and `I`).
The Actual Runs is equal to the Expected Runs for 1 player (`H`).

With the limited data provide, we are suggesting that 3 Players, `B`, `C` and `E` have the possiblity of exhibiting Hot Hands, since their actual runs are less than the expected nunmber of runs.

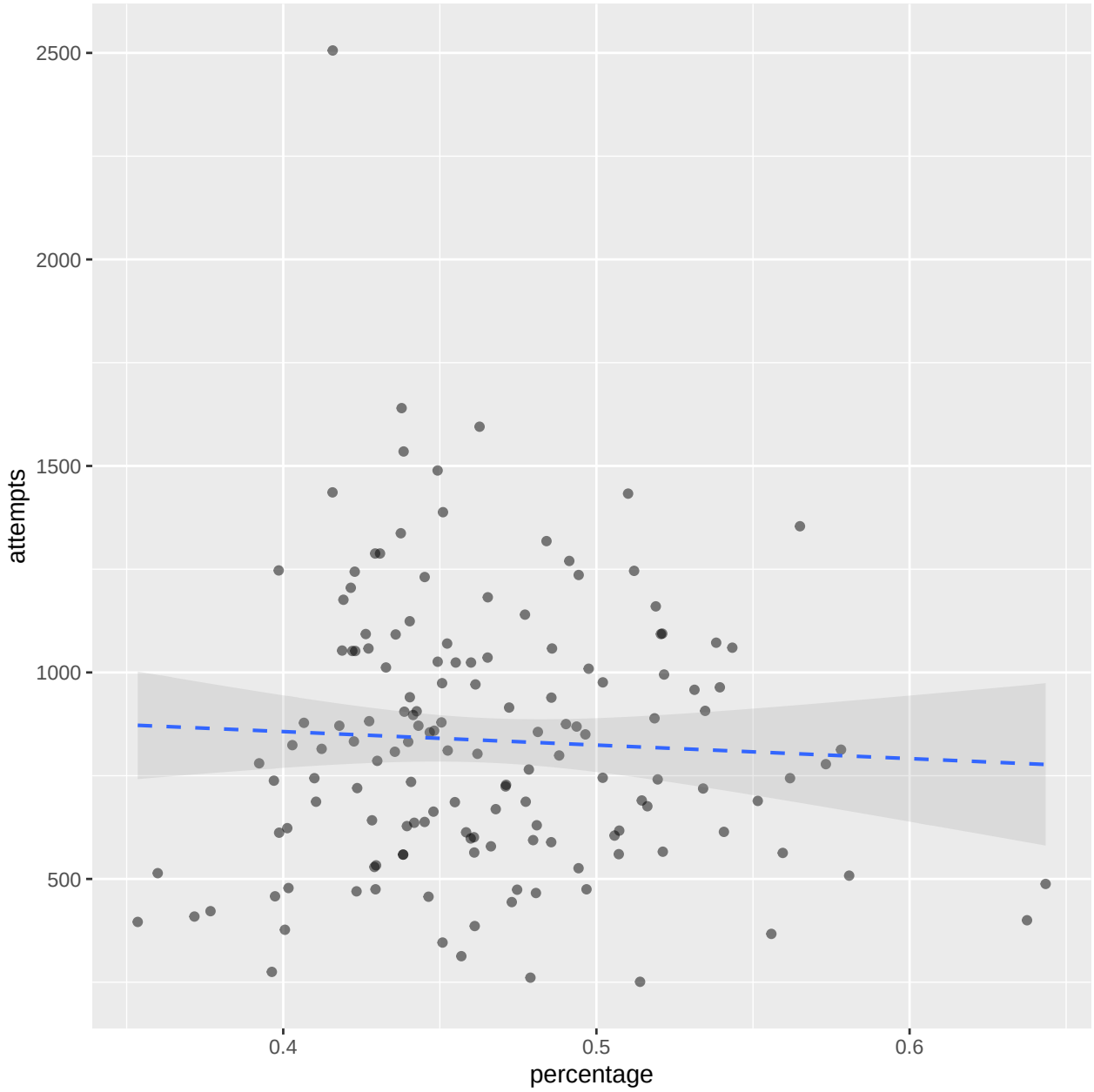
With such a small data set, the number of players that fall into each category are so closely related that we do not have sufficient evidence to conclusively support neither the existence nor the absence of the Hot Hands phenomenon.

Critical Thinking Using the basketball dataset, which has a few performance statistics for NBA players who started more than half of the games of the 2012/2013 season, we visualized to examine `points` (points scored) against `percentage` (field goal percentage).



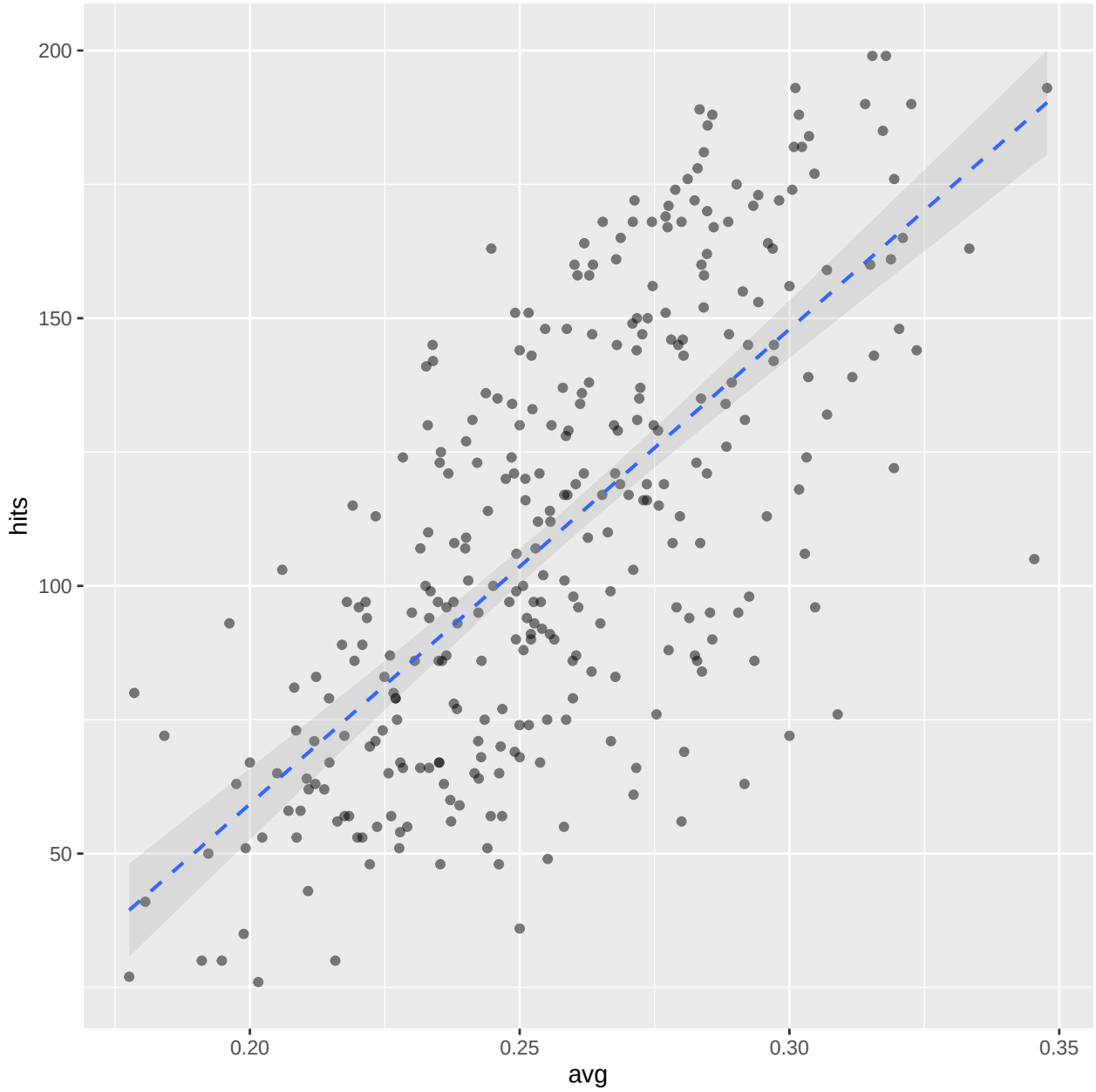
The data does not conform to a linear line of fit and there is no visual correlation between the points scored against the field goal percentage of this dataset.

Similarly, we examined `attempts`(field goal attemps) against `percentage`.



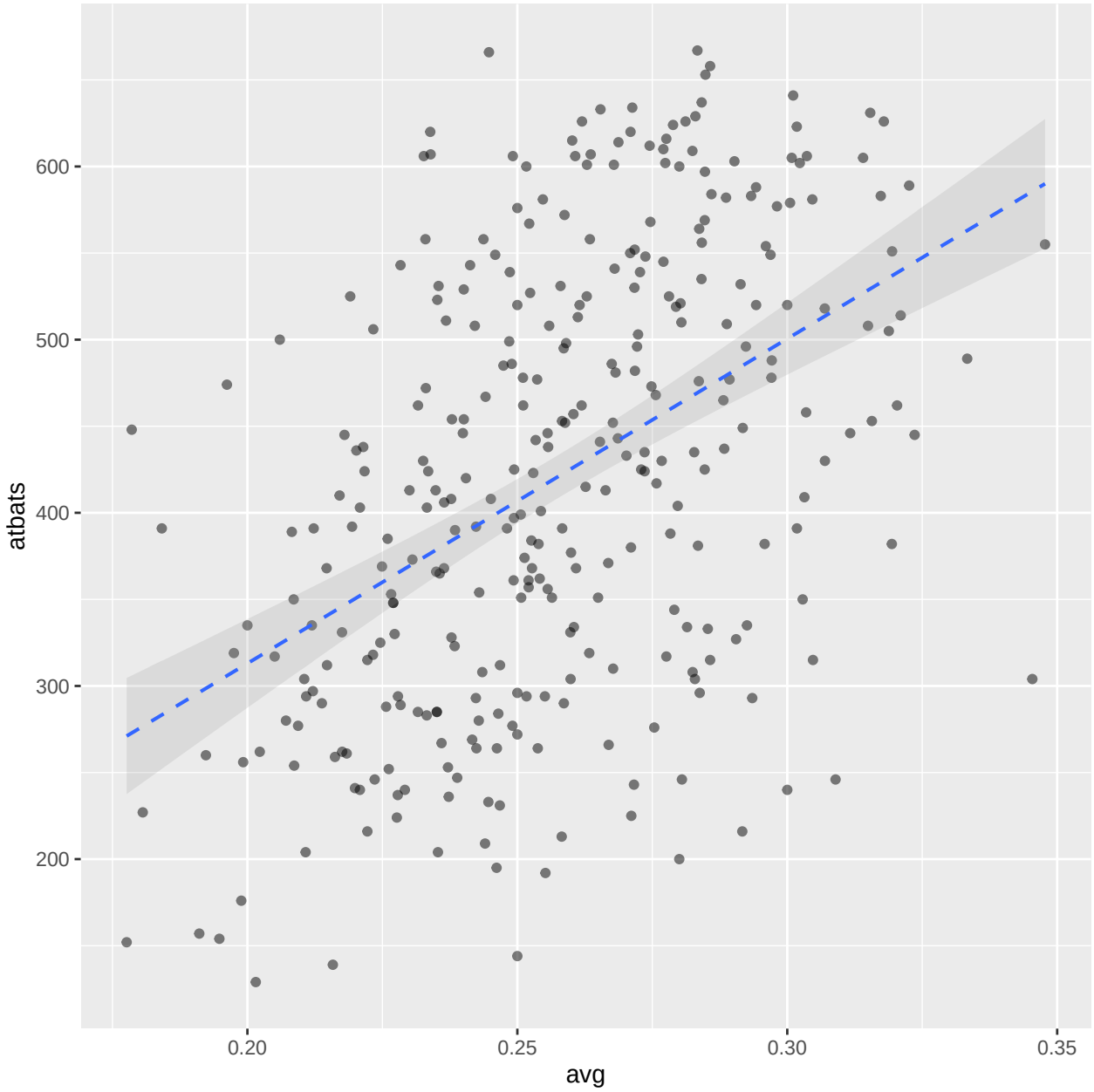
There is no correlation between the number of field goal attempts and the percentage in this data either.

We then used the baseball dataset, which has a few performance statistics for MLB players who started more than half of the games of the 2013 season. We used visualization to examine `hits` against `avg`(batting average).



When looking at the baseball data, there is a linear correlation between the hits and batting average. This can logically be reasoned as well because as the number of hits increases, the batting average also increases.

Similarly, we examined `atbats` (number of plate appearances) against `avg`.



Looking at these two factors, although there is a seemingly linear relationship between the batting average and the number of plate appearances, there are a lot more outliers to confidently claim linearity. This could be explained by the fact that although the number of plate appearances could be high, the chances that the batter has a high hit rate each time is unlikely. In other terms, although a player goes up to the plate 500 times, they could strike out half of those times and only the remaining portion, they would actually have a score to contribute to their batting average.

Discussion

1) The nature of the game would make Hot Hand easier to observe. In baseball you have a sequence of how the game is played, the game has a batting order and there are less variables to consider compared to other sports like basketball. You know how the game is going to be played, with the greatest variation existing in the type of pitch and the speed of the ball and which batters are going to be at the batters' box. The game is much slower than basketball as well. Batters are usually in the box alone against

the pitcher, without any external factors. The distance from the pitcher to the batter is also always consistently the same.

Basketball is much less structured; you do not know how the game is going to be played and the outcomes are also unknown. The only aspect that the players know are the rules. The level of difficulty of each shot is also unknown, players don't necessarily know beforehand the distance of the shots or if they are going to be guarded by a defender. Players are also much more prone to be influenced by other teammates' actions. This would make it much more difficult since the uncertainties that affect any regression model are much higher. With the influx of these variables, it becomes extremely difficult to accurately model and observe the Hot Hand phenomenon.

2) To establish that the Hot Hand exists in basketball, we would need a time series data set with minute-by-minute data on a large sample of NBA games from multiple seasons. Since the number of external factors are also much greater in basketball, we would need to account for all of those when considering a regression model.

3) To try and establish that Hot Hand exists, we would suggest the use of logistic regression, regressing a binary variable on multiple independent variables. We would then need to set a threshold to decide if the estimated value demonstrates that there is in fact Hot Hand evidence.

4) Critical predictor variables would include the type of shot (free throw, 2-pointer, 3-pointer). In addition to the minute-by-minute data requirement discussed above, it would also be important to know whether there was a defender in the picture. It might also be important to consider where the player is on the court relative to the hoop when actually making the shot. These predictive factors would help build a regressive model.

5) To begin this, we would require as much data as possible from multiple seasons to get a better and unskewed idea of shots.