Predicting the Spread Itself

James Whedee

October 18, 2015

Describing Spread

Here we have the distribution of spreads, with negative indicating that the home team is the favorite and positive indicating that the away team is the favorite. Data is from weeks 4, 6, 8, 9, and 16 of the 2014 season and weeks 2-6 of the 2015 season.

From this we can see the following trends:

Home vs. Away

- 1 in 3 spreads favor the away team
- 2 in 3 spreads favor the home team
- 1 in 50 spreads are picks

Most common spreads

- 1 in 4 spreads will be 3 for either team
- 1 in 2 spreads will be: -1, -2, -3, -4, -6.5, -7, or 3
- 2 in 3 spreads will be: -1, -2, -3, -3.5, -4, -6, -6.5, -7, -9, -10, 3, or 7

Away Team Favorites

- 2 in 5 spreads favoring the away team are 3 point spreads
- 1 in 2 spreads favoring the away team are 3 or 7 point spreads
- The third most common spread favoring the away team is 5.5

Home Team Favorites

- It is more likely to be a 1 or 2 point spread favoring the home team than 6.5 or 7
- The rarest spreads under 10 are -1.5, -4.5, -5 ,-5.5, -8, -8.5, -9.5

Spread Distribution
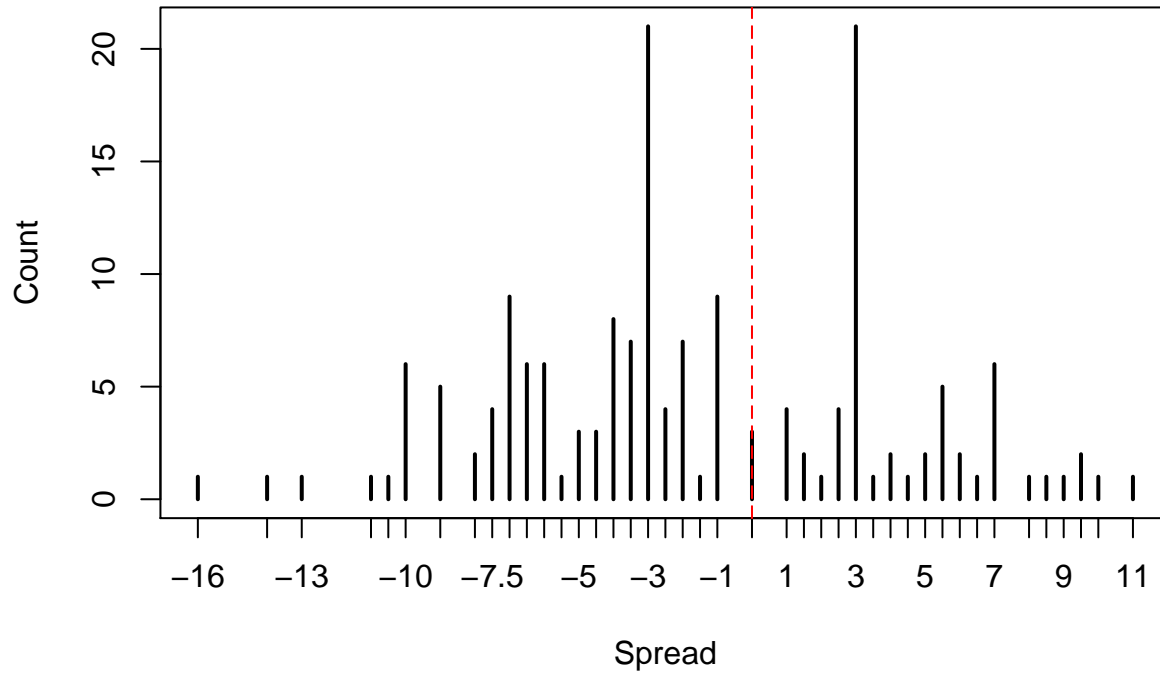

```
##
##          -3            3           -7           -1           -4         -3.5
## 0.12426036  0.12426036  0.05325444  0.05325444  0.04733728  0.04142012
##          -2          -10         -6.5           -6            7           -9
## 0.04142012  0.03550296  0.03550296  0.03550296  0.03550296  0.02958580
##         5.5         -7.5         -2.5            1          2.5           -5
## 0.02958580  0.02366864  0.02366864  0.02366864  0.02366864  0.01775148
##        -4.5            0           -8          1.5            4            5
## 0.01775148  0.01775148  0.01183432  0.01183432  0.01183432  0.01183432
##           6          9.5          -16          -14          -13          -11
## 0.01183432  0.01183432  0.00591716  0.00591716  0.00591716  0.00591716
##       -10.5         -5.5         -1.5            2          3.5          4.5
## 0.00591716  0.00591716  0.00591716  0.00591716  0.00591716  0.00591716
##         6.5            8          8.5            9           10           11
## 0.00591716  0.00591716  0.00591716  0.00591716  0.00591716  0.00591716
```

Exploratory Analysis

The following is exploratory analysis on predicting the final spread of NFL games using statistics from Football Outsiders.

As it turns out, DAVE is a very strong predictor of the Vegas spread. DAVE is a formula that combines pre-season rankings with regular season performance so the fact that it is correlated with the spread is unsurprising.
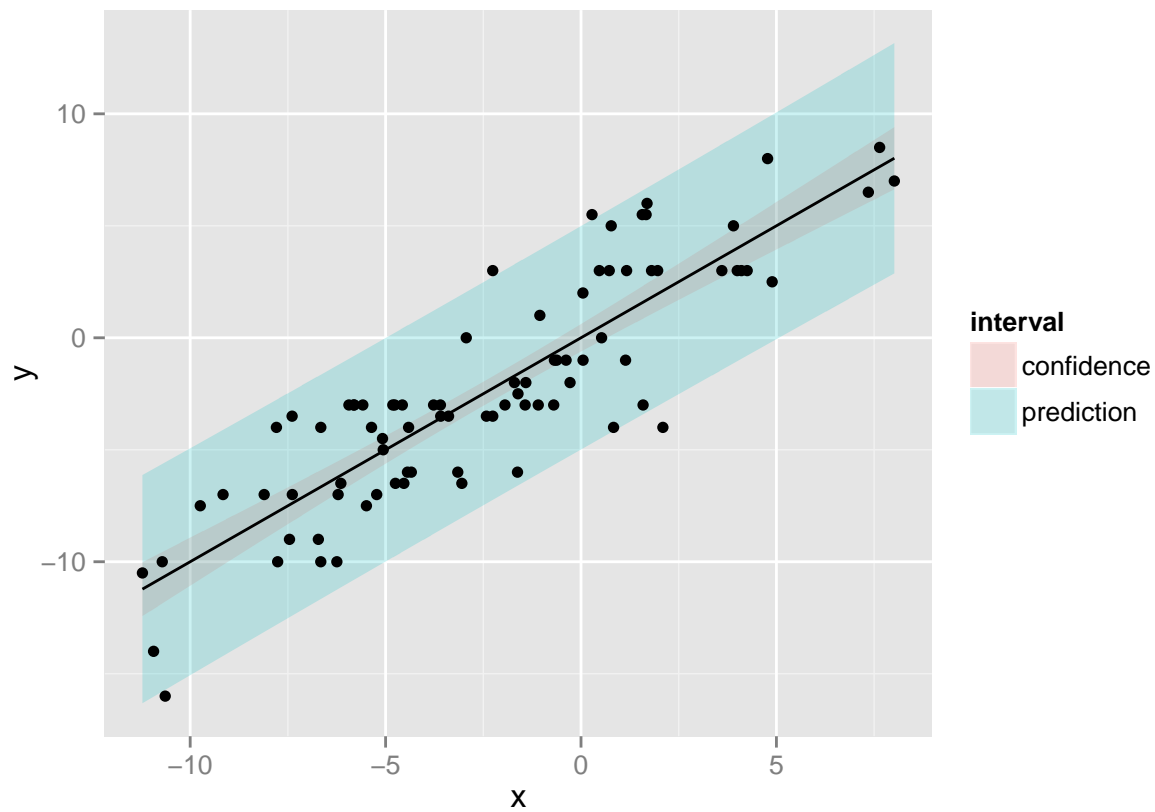
Residuals are normally distibuted, with SD of 2.6. Plot below shows proportion of correct guesses given a variable +-.
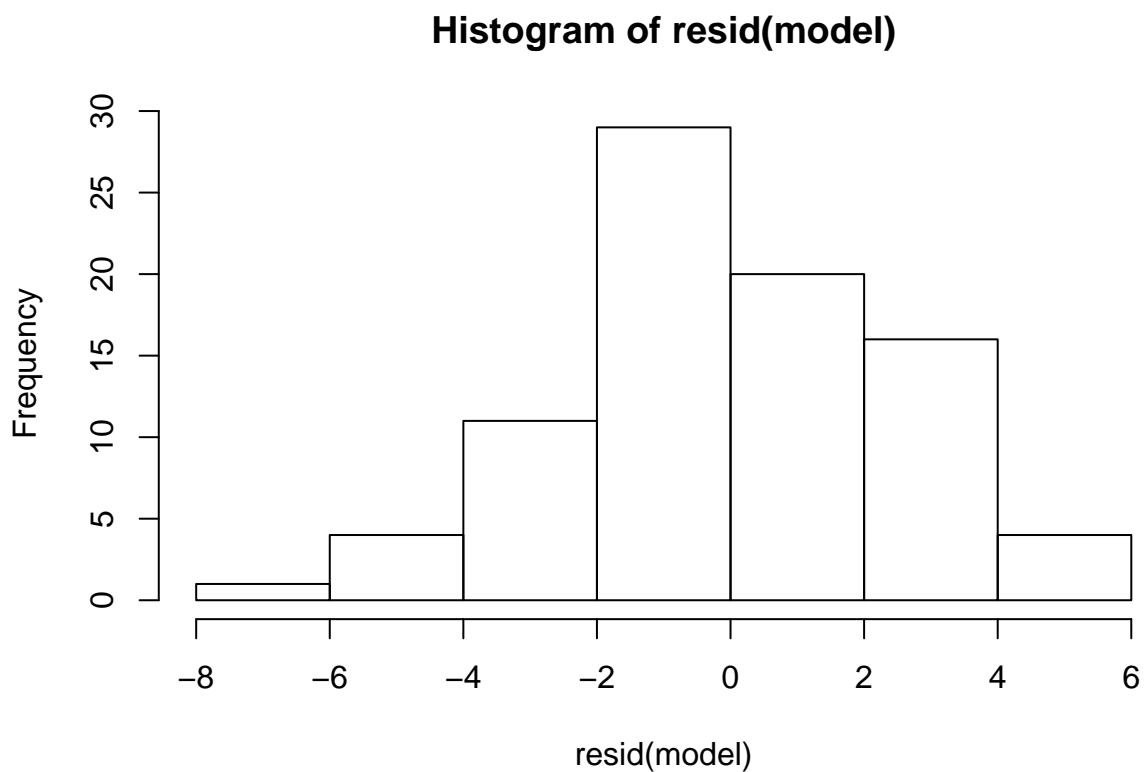
```
plot(resid(model))
```
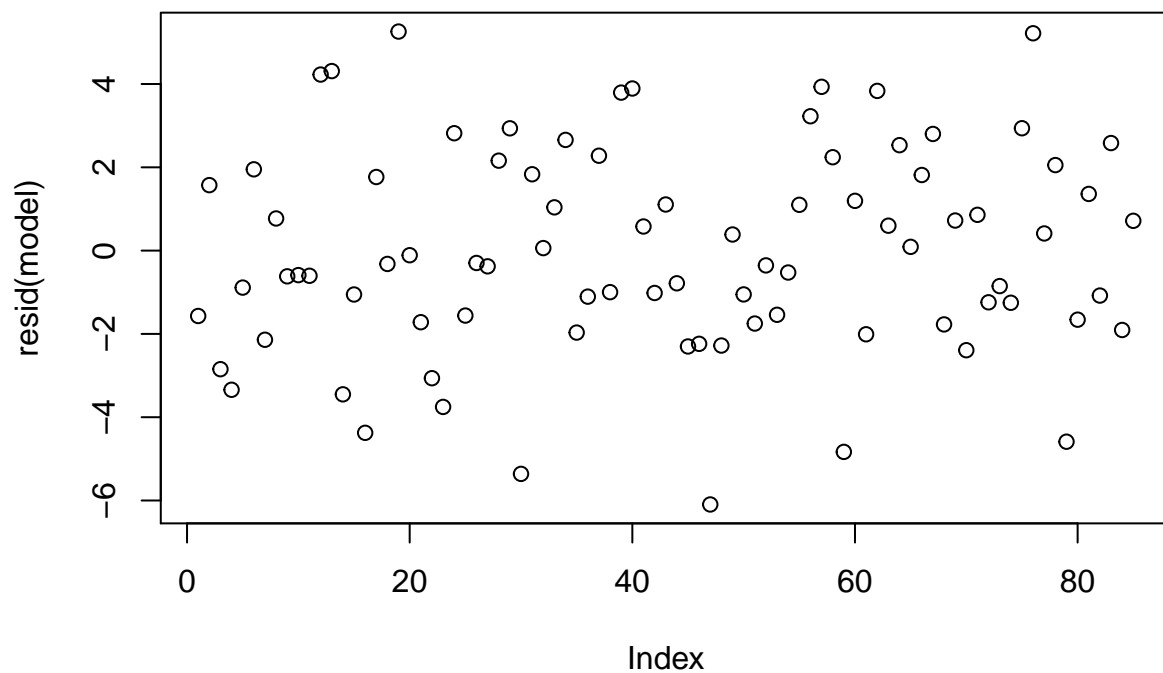
```r
quantile(abs(resid(model)))
```

```
##          0%         25%         50%         75%        100%
## 0.06048858 0.88754138 1.76703057 2.81860928 6.09489754
```
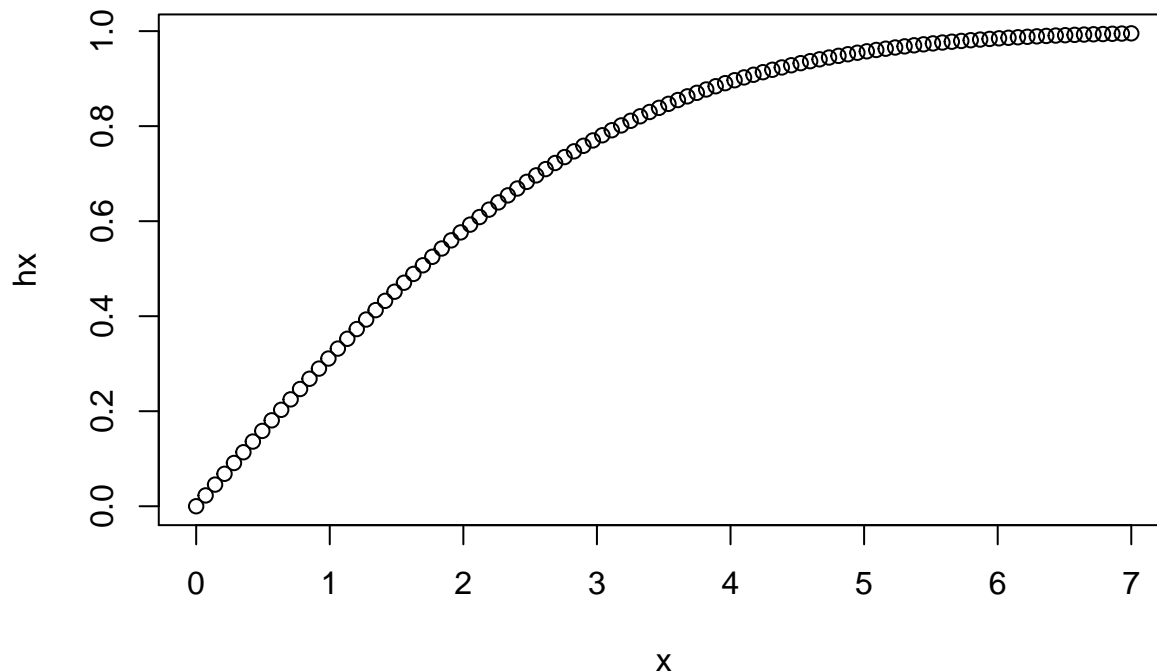
```r
hist(resid(model))
```

Histogram of resid(model)


```r
sd(resid(model))
```

```
## [1] 2.474394
```

```r
mean(abs(resid(model)))
```

```
## [1] 2.014802
```

```r
x <- seq(0,7,length=100)
hx <- pnorm(0+x,0,sd(resid(model)))-pnorm(0-x,0,sd(resid(model)))
plot(x,hx)
```

Removing outliers from injuries

Removing the main outliers, which are all a result of QB injury reduces the sd by quite a bit.

```
qbInjuries <- c(34,39,46,51,62,64,73,74,77,80,82,93,95,98,103)
cleanWeekPrime <- cleanWeekFrame[!(rownames(cleanWeekFrame) %in% qbInjuries),]
modelPrime <- lm("Spread~HDAVE+VDAVE",data=cleanWeekPrime)
summary(modelPrime)
```

```
##
## Call:
## lm(formula = "Spread~HDAVE+VDAVE", data = cleanWeekPrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3069 -1.4206 -0.5133  1.3860  5.0484
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.64794    0.25591  -10.35 1.58e-15 ***
## HDAVE       -0.23822    0.01715  -13.89  < 2e-16 ***
## VDAVE        0.19177    0.01875   10.23 2.58e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.119 on 67 degrees of freedom
## Multiple R-squared:  0.8048, Adjusted R-squared:  0.799
## F-statistic: 138.2 on 2 and 67 DF,  p-value: < 2.2e-16
```
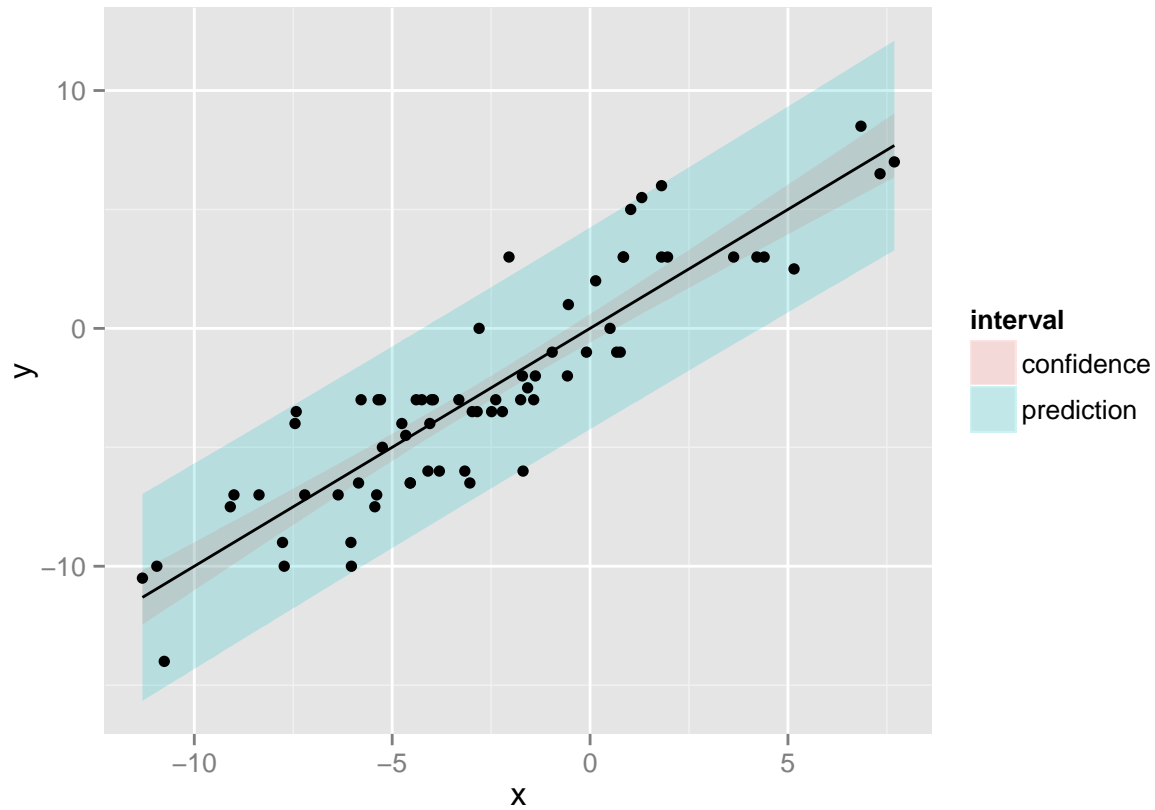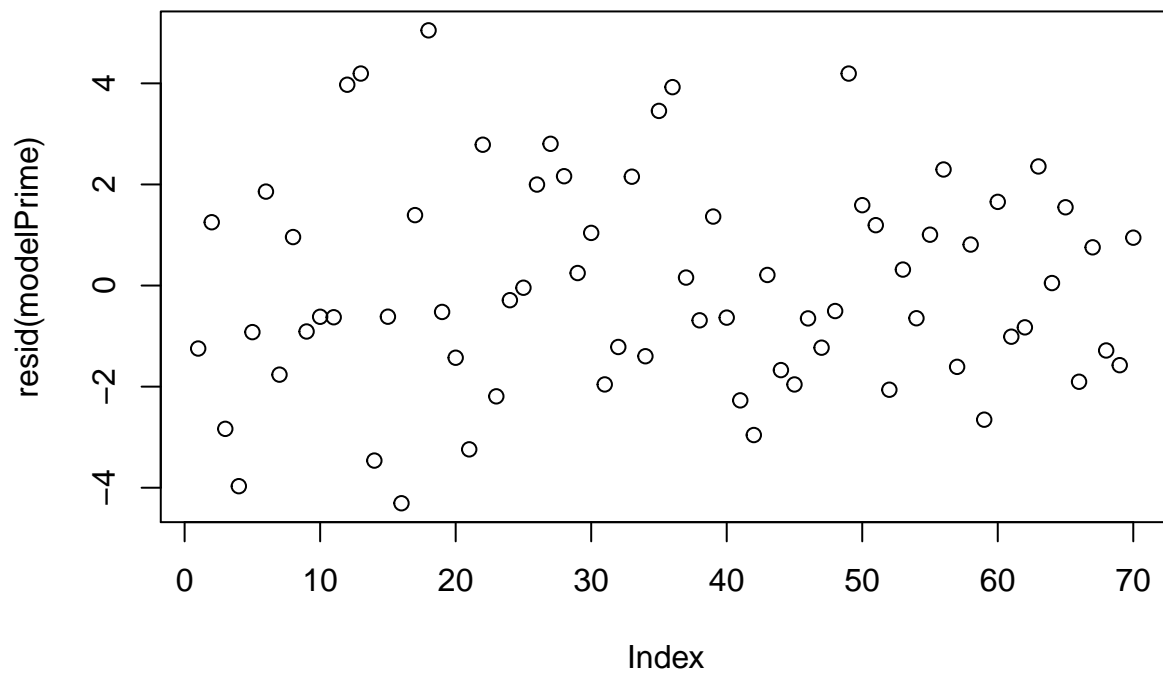
```
spreadFramePrime <- mutate(cleanWeekPrime,projSpread=coef(modelPrime)[1]+coef(modelPrime)[2]*cleanWeekP
plotConfandPredict(spreadFramePrime$projSpread,spreadFramePrime$Spread)
```


Here is the new residual analysis.
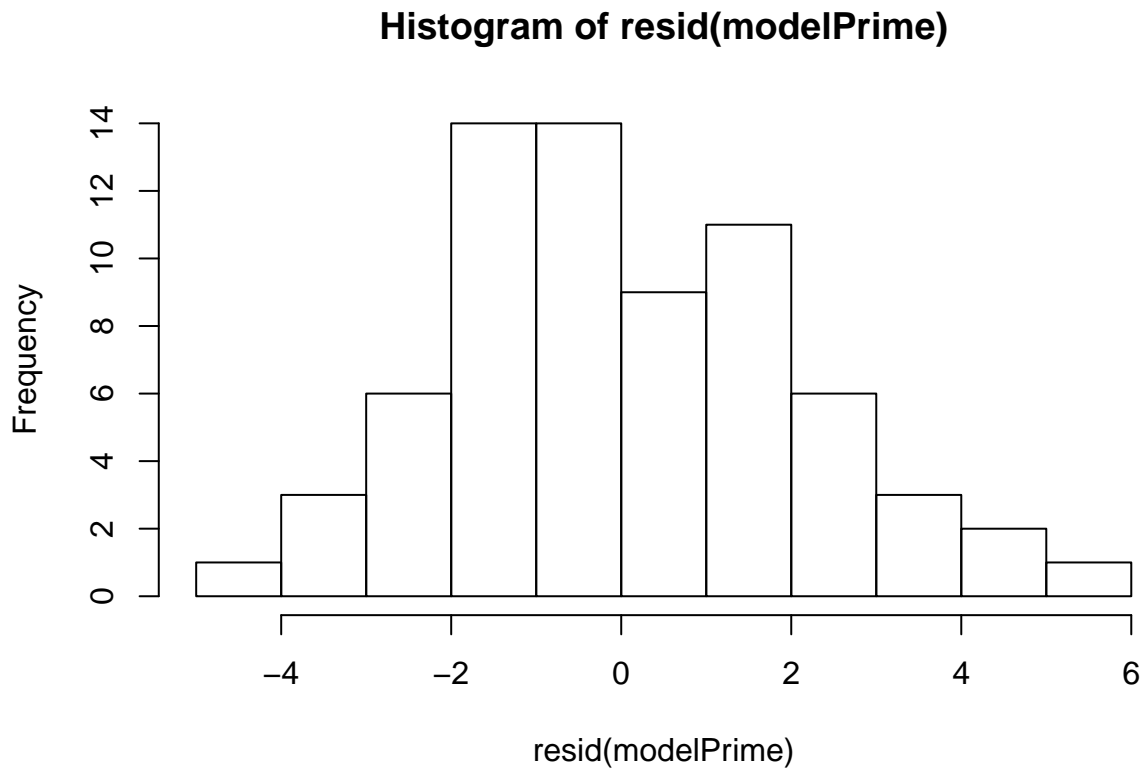
```
plot(resid(modelPrime))
```

```
quantile(abs(resid(modelPrime)))
```

```
##          0%         25%         50%         75%        100%
## 0.04331154 0.81483891 1.41334349 2.25201177 5.04844147
```

```
hist(resid(modelPrime))
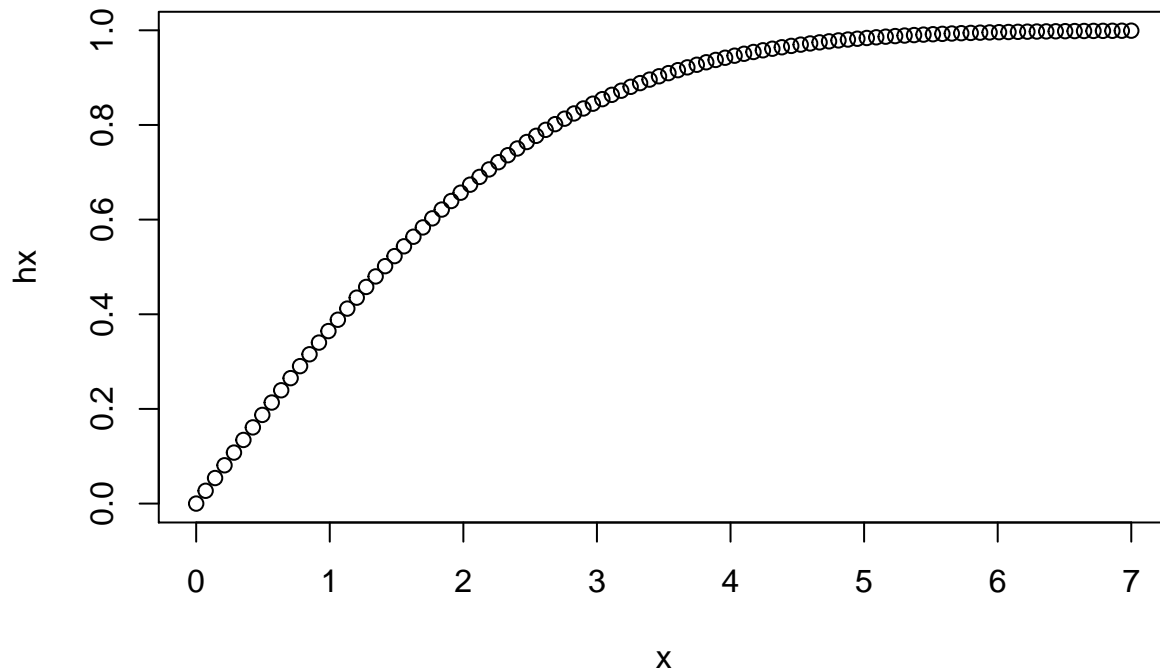```

## Histogram of resid(modelPrime)



```
sd(resid(modelPrime))
```

```
## [1] 2.088555
```

```
mean(abs(resid(modelPrime)))
```

```
## [1] 1.705735
```

```
x <- seq(0,7,length=100)
hx <- pnorm(0+x,0,sd(resid(modelPrime)))-pnorm(0-x,0,sd(resid(modelPrime)))
plot(x,hx)
```

Adjusting for magnetic points 3/7

Now, we will translate anything within 1-4 to 3 and within 6-9 to 7. The range from 4-6 will stay to hedge whether the line moves to 3 or 7. This means the residuals will no longer have normal distribution or mean of 0, but it will allow many more exactly correct projections.

```
quantile(abs(spreadFramePrime$projSpreadRES))
```

```
##          0%        25%        50%        75%       100%
## 0.04331154 0.81483891 1.41334349 2.25201177 5.04844147
```

```
quantile(abs(spreadFramePrime$projSpreadADJRES))
```

```
##   0%  25%  50%  75% 100%
##  0.0  0.5  1.5  2.0  6.0
```

```
sd(spreadFramePrime$projSpreadRES)
```

```
## [1] 2.088555
```

```
sd(spreadFramePrime$projSpreadADJRES)
```

```
## [1] 1.913786
```

```
mean(abs(spreadFramePrime$projSpreadRES))
```
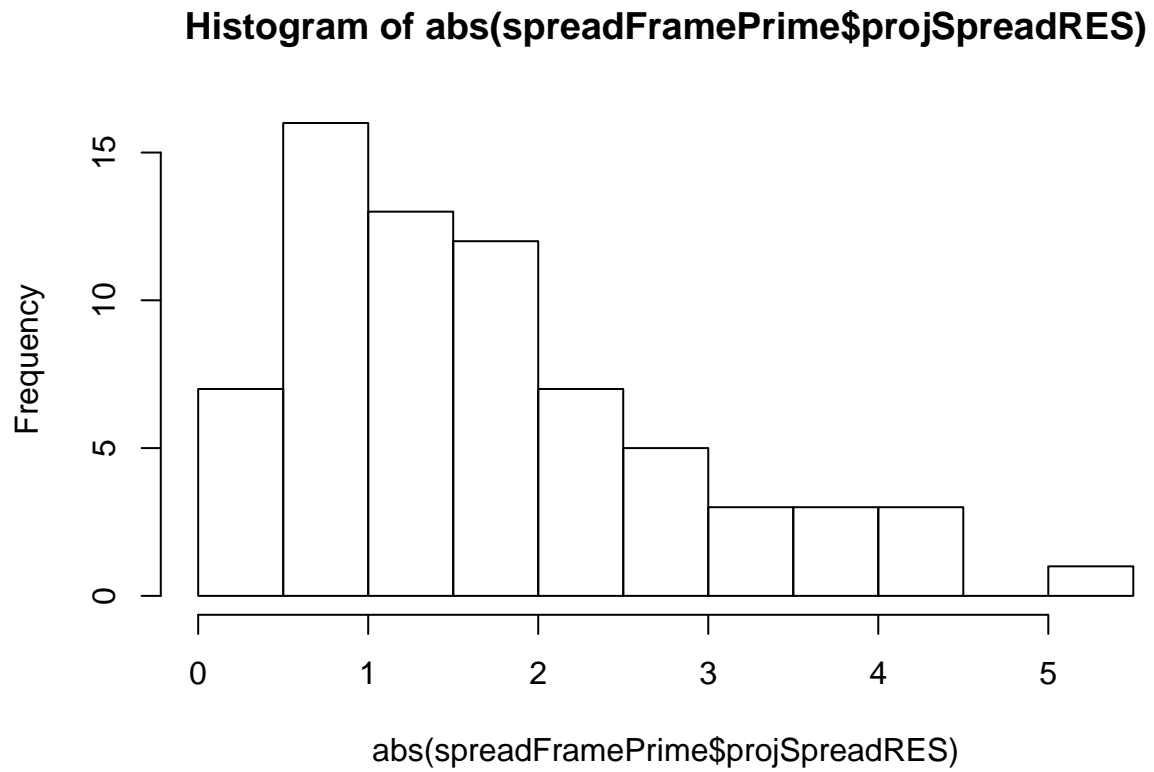
```
## [1] 1.705735
```

```r
mean(abs(spreadFramePrime$projSpreadADJRES))
```

```
## [1] 1.45
```

```r
hist(abs(spreadFramePrime$projSpreadRES))
```

Histogram of abs(spreadFramePrime$projSpreadRES)


```r
hist(abs(spreadFramePrime$projSpreadADJRES))
```

Histogram of abs(spreadFramePrime$projSpreadADJRES)