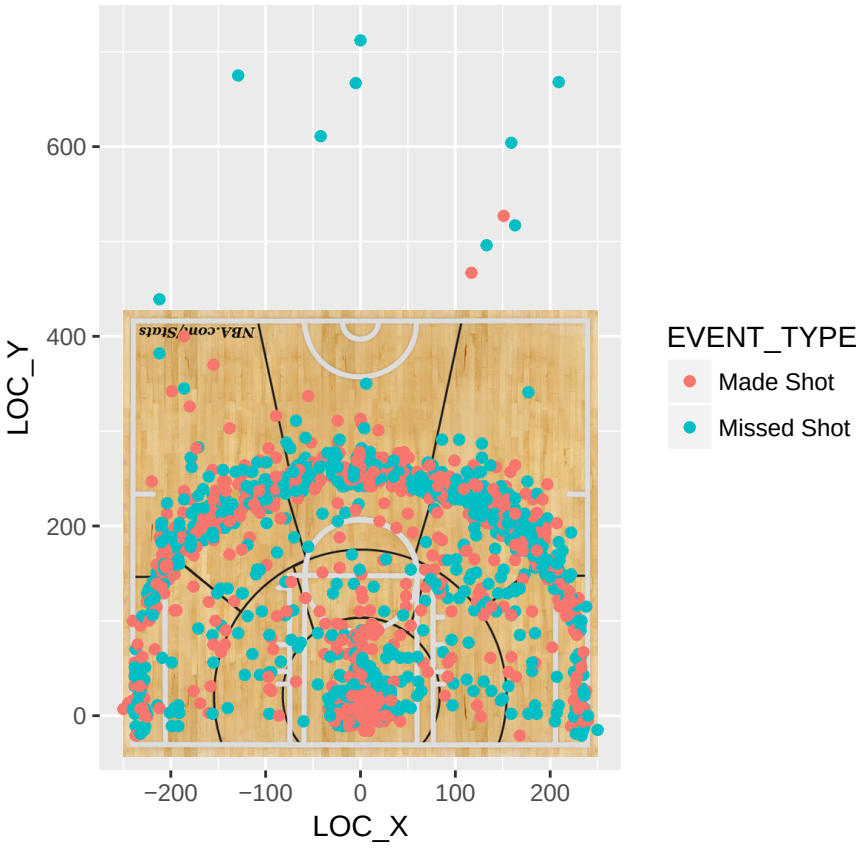
S670 Problem Set 7

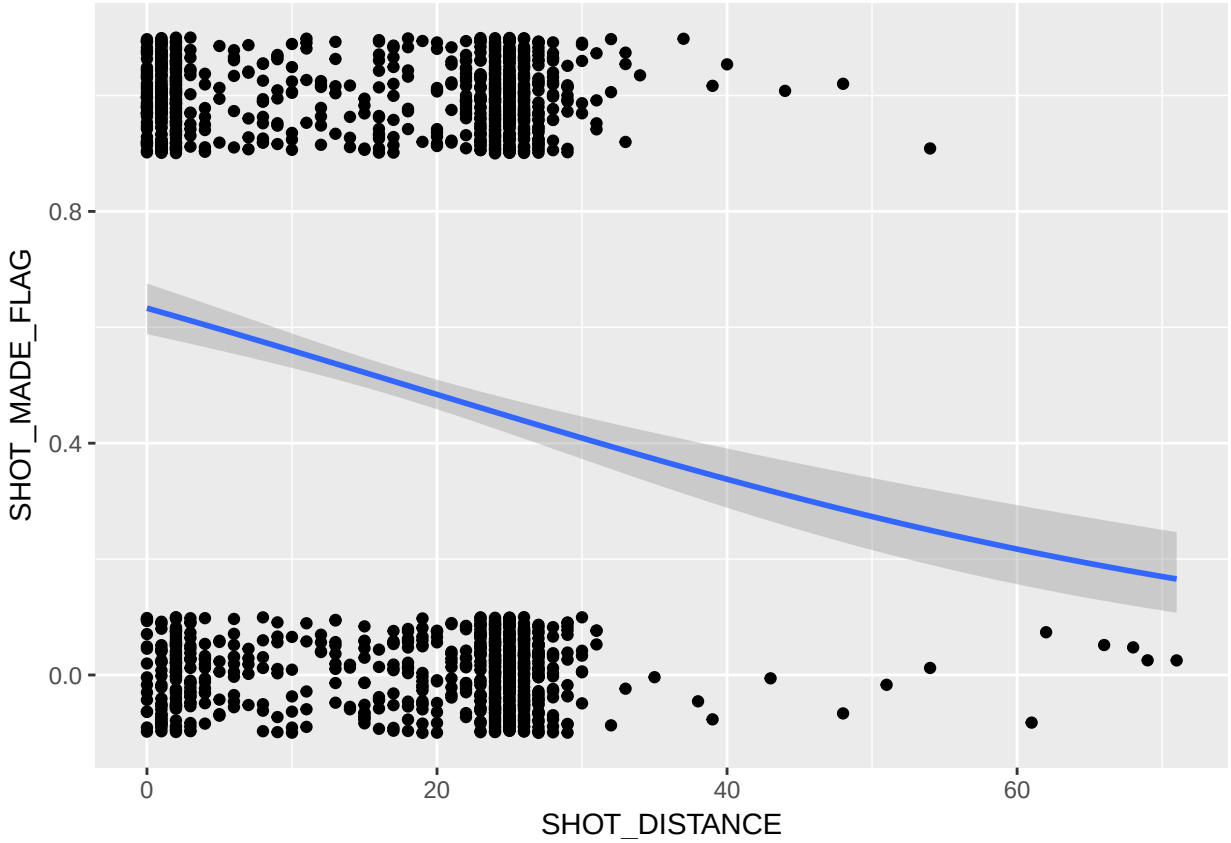
Erik Parker

March 28, 2017

1. Plot shots made/missed on loc_x vs. loc_y grid



2. Fit a logistic regression to predict whether the shot is made, using the single predictor SHOT_DISTANCE. Draw an appropriate ggplot of the fitted curve and write an equation for the fit.

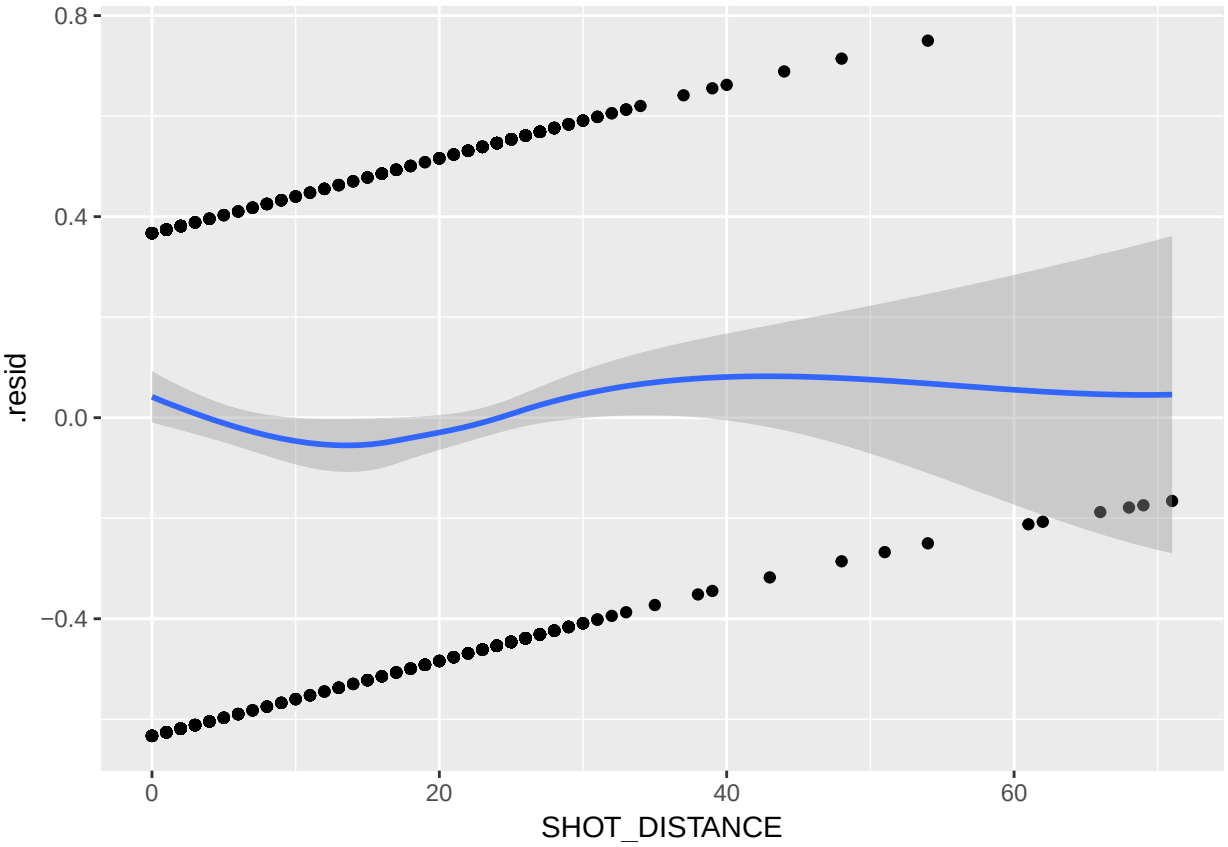



```
##
## Call:
## glm(formula = SHOT_MADE_FLAG ~ SHOT_DISTANCE, family = binomial,
##     data = curry)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4159  -1.0996   0.9563   1.2309   1.6654
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.54508    0.09606   5.674 1.39e-08 ***
## SHOT_DISTANCE -0.03045    0.00467  -6.521 6.97e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2215.2  on 1597  degrees of freedom
## Residual deviance: 2171.0  on 1596  degrees of freedom
## AIC: 2175
```

```
##
## Number of Fisher Scoring iterations: 4
```

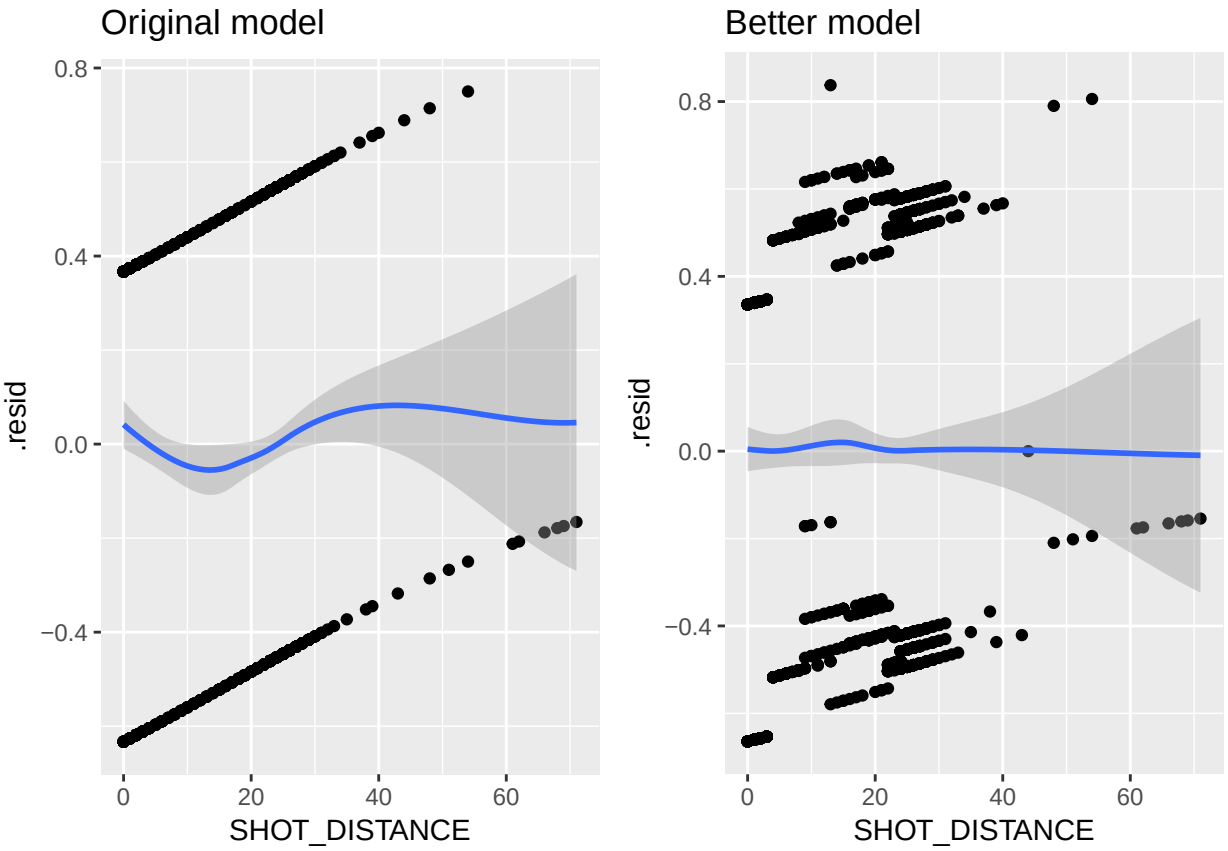
Equation for fit: logit(P(shot made)) = 0.54508 - 0.03045 * Shot distance

3. Plot the residuals in a way that shows where the logistic regression doesn't fit the data well. Describe in some detail how the model is inaccurate.



From viewing the plot of shot distance vs. the residuals, at the extreme left side (short distances from the basket) the average residuals are positive meaning that the model is underestimating Steph's ability to score in the paint. After that, as we move into the midrange, the average residuals fall into the negative range, meaning that the model is overestimating his shooting ability from about 5-23 feet. After 23 feet or so, as we continue to move away from the basket, we can see that the average residuals become quite positive and remain there, meaning that the model once again underestimates Chef Curry's accuracy - this time at long range.

4. Fit a better model. You could try a different functional form or a model with more predictors (as long as you use the predictors sensibly.) Your model doesn't have to be perfect, just better. Draw a graph that shows how your model differs from the simple logistic regression, and convince us that your model is better.



The residuals in this model are much better behaved than the previous model with the single shot distance predictor. Adding in the interaction between the two shot zone variables (as there is most certainly an interaction between them in reality) really improved the fit and leads to the model accurately predicting Curry's shot percentage at almost all ranges. The only real remaining problem is in the 10-18ft range where the model still underestimates Steph's probability of making the shot. But, overall, I am quite happy with my new model.