David Smith
STAT 4630
Homework 1

1.

a. $Var(X) = E[(X – E(X))^2]$ by the definition of variance, which equals $E[X^2 – 2XE(X) + (E(X))^2] = E(X^2) – 2E[XE(X)] + E[(E(X))^2] = E(X^2) – 2E(X)E(X) + (E(X))^2$ since E(X) is a constant, which leads to $E(X^2) – 2(E(X))^2 + (E(X))^2 = E(X^2) – (E(X))^2$.

b. $E[(Y - \hat{f}(x))^2] = E[Y^2 – 2Y\hat{f}(x) + \hat{f}^2(x)] = E(Y^2) – 2E(Y\hat{f}(x)) + E(\hat{f}^2(x)) = E(Y^2) – 2E(Y)E(\hat{f}(x)) + E(\hat{f}^2(x))$ since Y and $\hat{f}(x)$ are independent, which becomes $E(Y^2) – 2f^*(x)E(\hat{f}(x)) + E(\hat{f}^2(x)) = Var(Y) + (E(Y))^2 – 2f^*(x)E(\hat{f}(x)) + Var(\hat{f}(x)) + [E(\hat{f}(x))]^2$ by the formula in part a, which gives $E[(Y – E(Y))^2] + (E(Y))^2 – 2f^*(x)E(\hat{f}(x)) + E[(\hat{f}(x) – [E(\hat{f}(x))]^2] + [E(\hat{f}(x))]^2$ by the definition of variance, which equals $E[(Y – f^*(x))^2] + [f^*(x)]^2 – 2f^*(x)\bar{f}(x) + \bar{f}^2(x) + E[(\hat{f}(x) - \bar{f}(x))^2] = E[(Y – f^*(x))^2] + [f^*(x) - \bar{f}(x)]^2 + E[(\hat{f}(x) - \bar{f}(x))^2]$.

2.

See R file.

3.

b. Most of the charts show points that are fairly spread out and vaguely linear, but there are a few interesting exceptions. The relationship of the variable that counts the percentage of greens hit with each other variable shows points that are clustered tightly together. The relationship of the variable that counts the average amount of winnings per round with each other variable shows points that are clustered near 0 for each other variable, since Tot_winnings_per_round is so high compared to the other variables.

c. We should use Log_tot_win_per_round. For one, the pairs function shows that it has a stronger linear relationship with each other variable than most other variables have with each other. More importantly, Tot_winnings_per_round has much larger values than the other variables, as mentioned in part b, and as a result is more variable, skewing the data from the other variables. Running a regression on Log_tot_win_per_round would be much more useful.

d. Most of the coefficients of our model are very small, with the largest by far being the intercept, with a value of about 14.79. Our value of adjusted R^2 is 86.71%, which is quite good, showing that our model explains much of the variability of the response variable. The F-test shows that our model has overall significance, with a very small p-value. However, three of our variables are not significant when setting $\alpha = 0.05$. Driver_Distance has a p-value of 0.125, Fairway_Pct has a p-value of 0.335, and Tot_rounds has a p-value of 0.109. There p-values are not terribly large, but we might want to run a regression of several models omitting some of them, and compare.

e. The predictions are fairly accurate, not too far from the true values. Most of the predictions are between 7 and 9, while the actual data often goes above 9 or below 7. Thus, the predictions have less variability than the true values, tending towards the mean of the data.