Chase Joyner

885 Homework 3

December 11, 2015

Problem 1:

In this problem, we analyzed the drug data set. This data set consisted of the following columns of data:

1. Identification code
2. Age (x_1)
3. Depression score (x_2)
4. Drug use history (x_3): 1 = Never (h_1), 2 = Previous (h_2), 3 = Recent
5. Number of prior drug treatments (x_4)
6. Race (x_5): 0 = White, 1 = Otherwise (r_1)
7. Treatment randomization id (x_6): 0 = Short, 1 = Long (t_1)
8. Treatment site (x_7): 0 = Site A, 1 = Site B (s_1)
9. Response variable (Y) – Remained drug free for 12 months: 1 = Yes, 0 = No.

Before modeling the data, we discuss what our intuitions tells us that we should expect. Since the response variable is whether or not the patient remained drug free, we could expect depression score, drug use history, number of prior drug treatments, and treatment randomization to all have an effect on the response variable. Age, race, and location of treatment might not have much of an effect on the response variable and so we should expect these covariates to not be included in the final model. However, now we analyze the data set to see.

To begin the model building process, we considered a logit link and probit link since the response variable is binary. The glm function R under these two links returns the same conclusion: to keep x_1, h_1, x_4, and t_1 in the model, given everything else is included. The AIC for the logit link was 637.2 and the AIC for the probit link was 637.4. Next, I considered using the step AIC function in R to see if we can find another reasonable model. The step AIC for both of these link functions concluded to only use the predictors x_1, h_1, x_4, and t_1, which glm deemed significant. The AIC for the logit model with predictors x_1, h_1, x_4, and t_1 is 635.4, while the AIC under the probit model with these predictors is 635.7. Since the logit link continuously had smaller AIC, we proceeded with the logit link. Next, we used ANOVA to check if this reduced model is sufficient before proceeding.

Running the ANOVA function in R, we obtain a large p-value of 0.6721, indicating that this reduced model is sufficient. We proceeded with this reduced model, which is

$$\text{Model 1: } x_1, h_1, x_4, t_4.$$

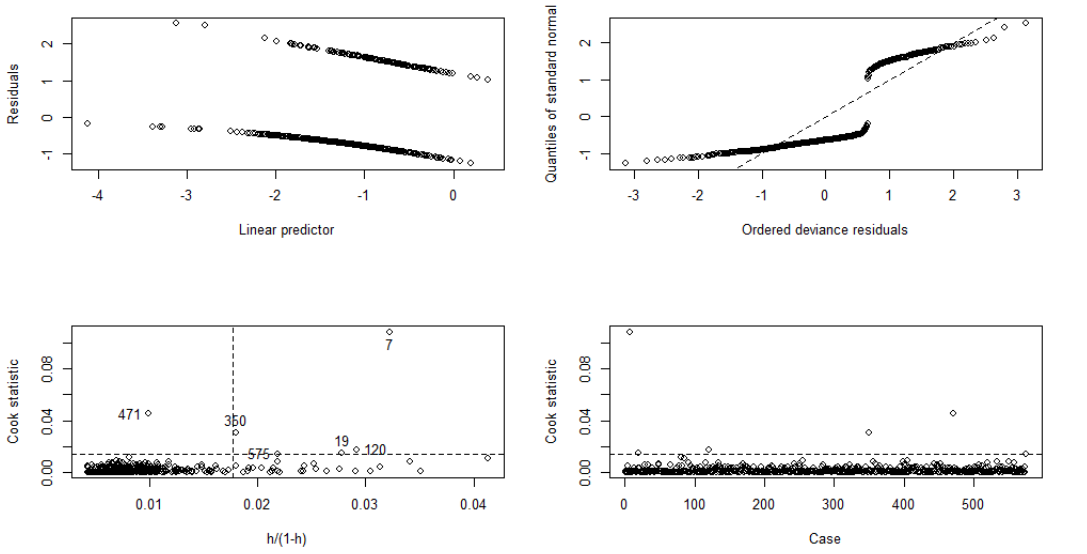
Next, we considered the possibility of interaction terms, and so we fit a model using the reduced model with the inclusion of all interactions terms. In doing so, we see that the interaction $x_1 : x_4$ might be significant, given everything else is in the model. We added this variable to the reduced model and saw that x_1 became insignificant, but AIC only dropped to 628.7. So, for parsimonious reasons, we dropped the interaction and kept x_1. Next, we checked for significant higher order terms. That is, we considered adding x_1^2 and x_4^2 to the reduced model. In doing this, we saw that neither of these terms were significant. Therefore, we still consider the reduced model given above.

The next step was to run bestglm for competing models. We had this function return the top 3 models, in which the best model according to bestglm was our current reduced model. The other two models were

$$\text{Model 2: } x_1, h_1, x_4, t_1, r_1$$
$$\text{Model 3: } x_1, h_1, t_1, s_1.$$

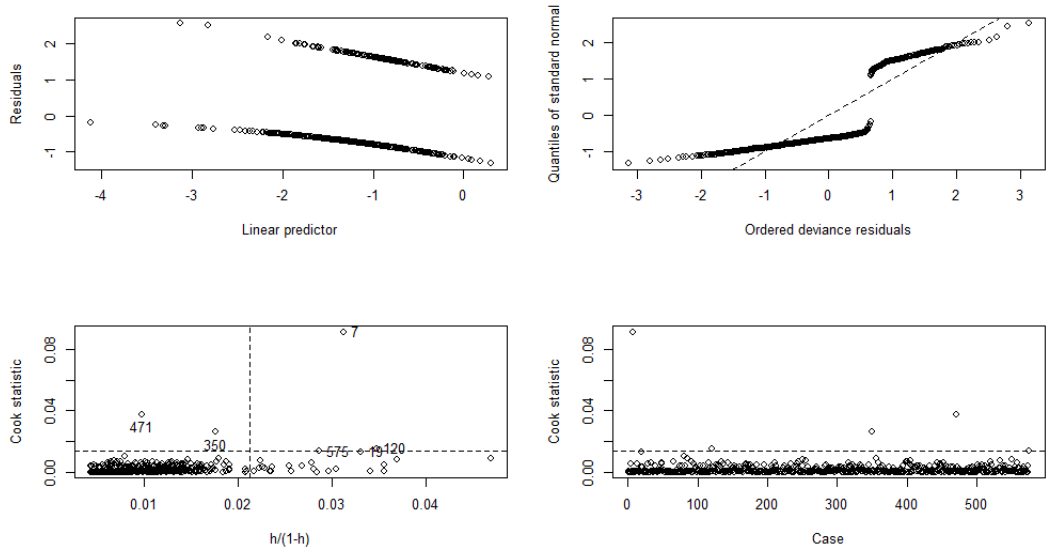
The AIC for model 2 according to bestglm was 630.0506 and the AIC for model 3 was 630.5289. It should be noted that the AIC for our reduced model according to bestglm was 628.9877, not 635.4.

Now we have three competing models. Before considering predictive power of these competing models, we considered plot diagnostics. The residuals plot for model 1 is shown below:



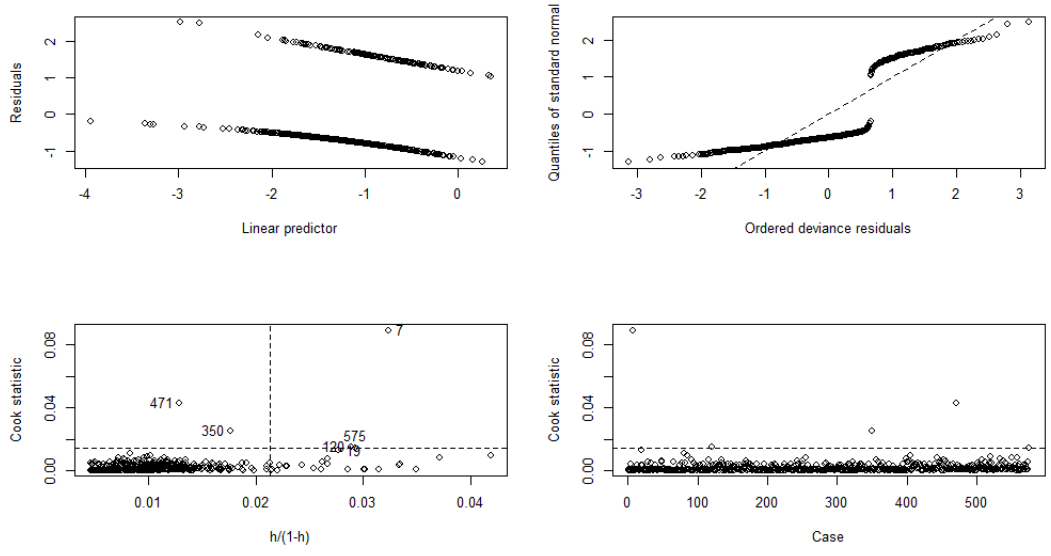
It can be seen that the residuals plot shows nothing out of the ordinary for a logit model. Also, it appears that observation 471 is influential and observations 350, 575, 19, 120, and 7 are all influential and high leverage. These point should be remembered in the predictive power analysis.

The residuals plot for model 2 is shown below:



.

Here, we see the only difference is that observation 350 became not high leverage. The last residuals plot is for model 3, shown below:



.

The only difference between this plot and the plot for model 1 is that it appears that observation 19 became not influential. We proceeded with cross-validation to consider the predictive power of these three models.

To check the predictive power of these three models, we randomly chose to remove observations 25, 78, 111, 165, 199, 208, 344, 407, 492, and 536. Then, we refit the three models and predicted

these observations. In doing this, the sum of the squared deviations from the true values are recorded below:

$$\text{res}_1 = 22.1413, \quad \text{res}_2 = 23.6422, \quad \text{res}_3 = 22.5605.$$

Therefore, since model 1 has the smallest prediction error, we chose this as our final model. The model is specified as

$$\begin{aligned}\text{logit}\left(\widehat{\pi}_i\right) &= -3.08509 + 0.053083x_{i1} + 0.735059h_{i1} - 0.066286x_{i4} + 0.466711t_{i1} \\ \widehat{\pi}_i &= \frac{\exp\{-3.08509 + 0.053083x_{i1} + 0.735059h_{i1} - 0.066286x_{i4} + 0.466711t_{i1}\}}{1 + \exp\{-3.08509 + 0.053083x_{i1} + 0.735059h_{i1} - 0.066286x_{i4} + 0.466711t_{i1}\}}.\end{aligned}$$

A shortcoming of this model includes the amount of observations that were high leverage or highly influential since this can greatly affect the estimates, but there is not much that we can do about this. Another shortcoming is that our intuition was not met and so it could be difficult to explain this model.

Problem 2:

For this problem, we analyzed the crab data set. The crab data set consisted of the following data:

1. ID
2. Crab color $(c_1), (c_2), (c_3)$
3. Spine conditions $(s_1), (s_2)$
4. Carapace width (x_3)
5. Weight (x_4)
6. Satellites (Y).

The response variable Y is the number of satellites (male crabs) attracted to a female crab. Again, before modeling the data, we discuss any intuitions that we might have. Perhaps the color of the crab and spine conditions won't have much effect on the number of satellites while the width and weight (essentially the size) might.

To begin the model building process, we considered a poisson with log link model since we have count data. The glm function in R under this link returns that x_4 and c_1 are the only significant predictors with an AIC of 920.86. However, in this situation, it does not make sense to collapse any of the factor levels. Next, we consider the step AIC function in R to identify other possible models. In doing so, we see that a possible reduced model is x_4, c_1, and c_2 as predictors with an AIC of 915.08. We quickly run an ANOVA test to test the adequacy of this model. This returned a p-value of 0.6959 and therefore the reduced model is sufficient. This model is

$$\text{Model: } x_4, c_1, c_2.$$

The next step is to consider interaction terms. We fit a model including all possible interaction terms of model 1 and saw that interactions $x_4 : c_1$ and $x_4 : c_2$ were significant, given everything else in the model being included. This gave a model with an AIC of 908.55, which seems to be a

considerable reduction from the AIC of model 1. Next, we considered adding higher order terms, specifically x_4^2. This model said that c_2 and $x_4 : c_2$ were insignificant, with an AIC of 902.55. However, perhaps $x_4 : c_2$ became insignificant since x_4^2 was included and c_2 became insignificant because of $x_4 : c_2$. Therefore, we considered dropping the interaction term $x_4 : c_2$ to see if c_2 became significant, which it did. This model had an AIC of 901.83 and became one of our models, i.e.

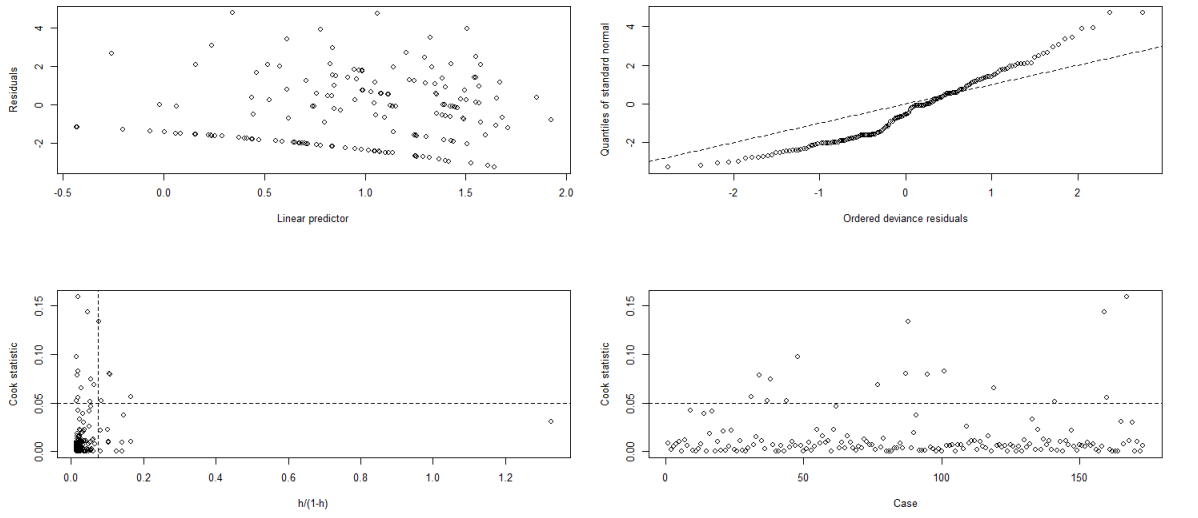
$$\text{Model: } x_4, c_1, c_2, x_4c_1, x_4^2.$$

The next step was to use bestglm in R to obtain other candidate models. This function returned that x_3 could be quadratically related to the response variable. We used the best two models returned by bestglm since our model 1 seems better than the third best model according to bestglm. Thus, the three candidate models are

$$\begin{aligned}
&\text{Model 1: } x_3, x_4, c_1, c_2, x_3^2 \\
&\text{Model 2: } x_4, c_1, c_2, x_4c_1, x_4^2 \\
&\text{Model 3: } x_3, x_4, c_3, x_3^2.
\end{aligned}$$

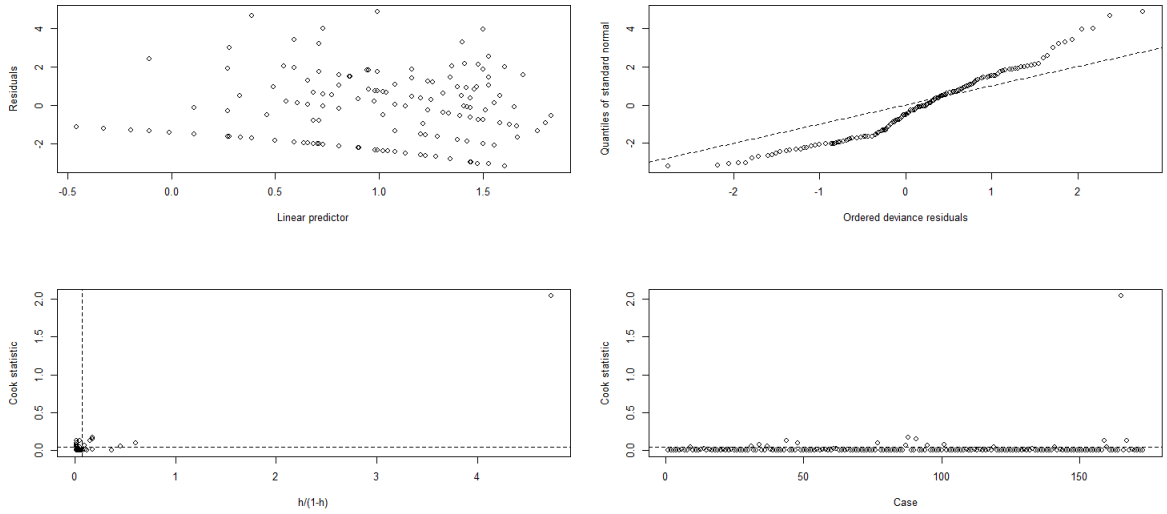
The AIC for model 1 according to bestglm was 901.078 and the AIC for model 3 was 901.906. Recall that the AIC for model 2 was 901.83.

Now we have three competing models and so we run diagnostics on these. The diagnostics plot for model 1 is shown below:



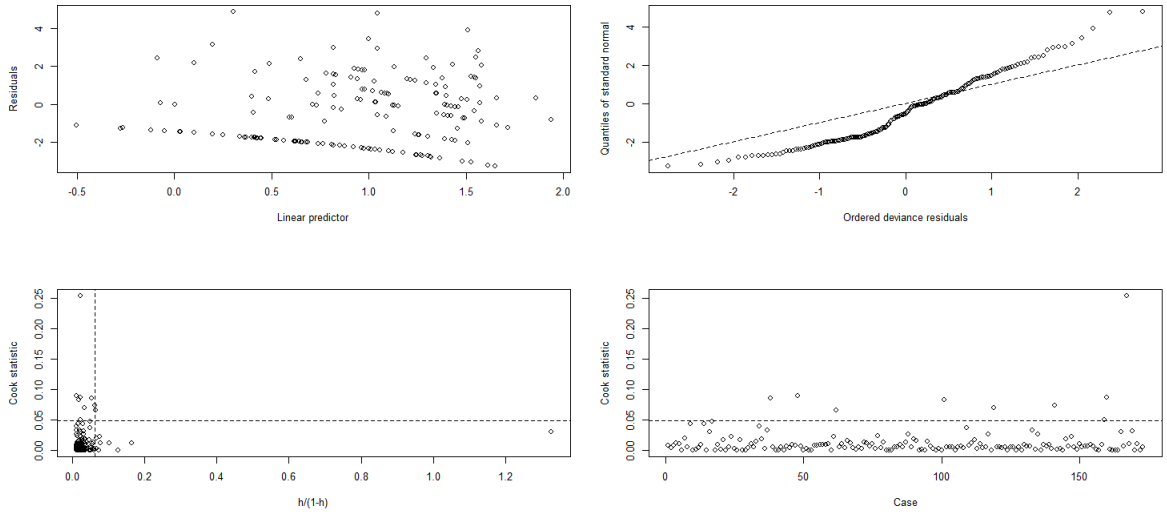
Looking at this plots, it should be noted the amount of observations which influence and leverage. It appears that about 15 observations have high influence on the model and about 16 with high leverage Also, we see that one observation has very high leverage. The plots for model 2 is shown

below:



.

Again, we see quite a few with high influence and high leverage. Specifically, about 15 influential observations and about 11 high leverage observations, with one observation have very high influence and leverage. The plots for model 3 is shown below:



.

Like the plots for models 1 and 2, we see about 9 observations with high influence and about 11 with high leverage, with one observation having extremely high influence and one with extremely high leverage. The last step was to cross-validate these three models to choose the model with the smallest prediction error.

To assess the predictive power of the three candidate models, we randomly removed 5 observations. The observations that were left out were 25, 63, 109, 150, and 173. The prediction error for the three models were

$$\text{res}_1 = 2.59, \quad \text{res}_2 = 2.85, \quad \text{res}_3 = 3.06.$$

Here, we see that model 1 has the smallest prediction error and the smallest AIC and therefore we chose this model as our final model. It can be specified as

$$\log(\widehat{\mu}_i) = -20.83352 + 1.49463x_{i3} + 0.678886x_{i4} + 0.378522c_{i1} + 0.234012c_{i2} - 0.027755x_{i3}^2.$$

A shortcoming of this model is the quadratic term x_3, which says that the carapace width is quadratically related to the number of satellites. This could be difficult in explaining our model. Also, it appears that crab color was significant, which goes against our intuition. However, perhaps an expert in this area might say that crab color is indeed important.

Problem 3:

The last data set to be analyzed is the LA marathon data. This data consisted of the following:

1. Sex (s_1)
2. Age (x_2)
3. Finish time (Y)
4. Place.

The response variable Y is the finishing time of the runner. We have given sex dummy variables, where s_1 represents females and males is baseline. It should be noted that we had 93 missing observations, in which we simply removed them since we have over 21500 observations. Before modeling the data, we intuitively expect both sex and age to have an effect on finishing time.

Since we had a continuous response variable, and we believe the data might be right skewed, we considered a gamma model with log or inverse link. We see both links claim sex is significant, but age is not. This goes against our intuition. It should also be noted that the log link gives females a positive coefficient, which makes sense. However, the inverse link gives females a negative coefficient, which does not make much sense. The log link had an AIC of 67296 and the AIC of the inverse link was also 67296. Next, we ran the step function in R to consider other models. The step function returned the same two models and so we next checked if the reduced model of just sex (s_1) was adequate. Running the ANOVA test in R, we obtain a p-value of 0.2459 for the log link and a p-value of .2753 for the inverse link, and so these reduced models are sufficient.

From here, we proceeded with just the inverse link since there does not appear to much difference and this is the canonical link. The next step was to consider higher order terms and interactions. We found that the interaction $s_1 : x_2$ is significant given s_1 and x_2 in the model, and including this in the model gave an AIC of 67293. We considered adding in the quadratic term x_2^2 given s_1 and x_2 in the model, and found it to be significant with an AIC of 65574. The next natural step was to try adding the interaction term and the quadratic term to the model given s_1 and x_2 in the model. We found that all terms were significant and this model had an AIC of 65569. This became our model 1, i.e.

$$\text{Model 1: } s_1, x_2, x_2^2, s_1x_2.$$

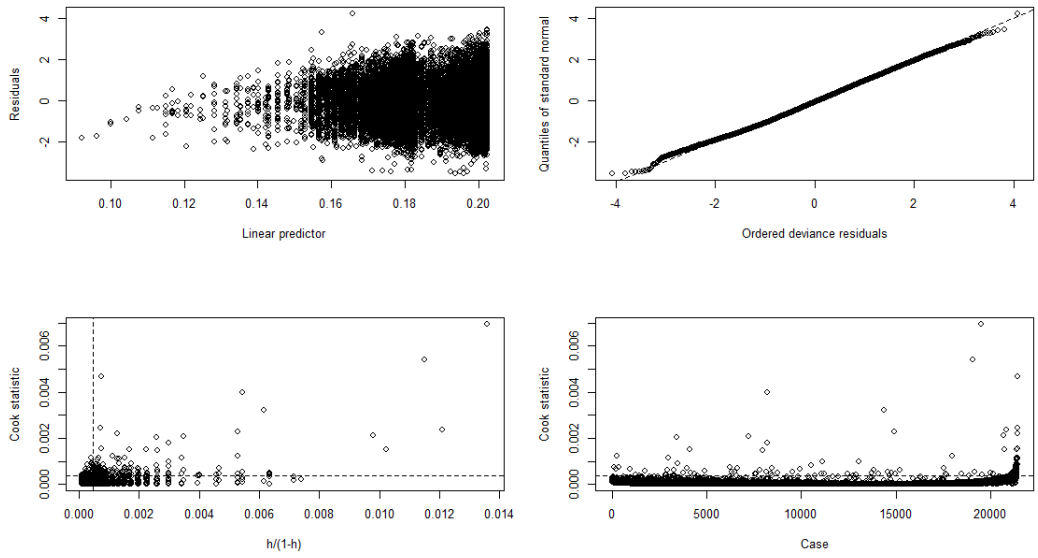
Next, we ran the bestglm function in R to obtain other candidate models. In doing so, we see that model 1 is the best model returned by this function with an AIC of 65564.57, not 65569. The

next two models are

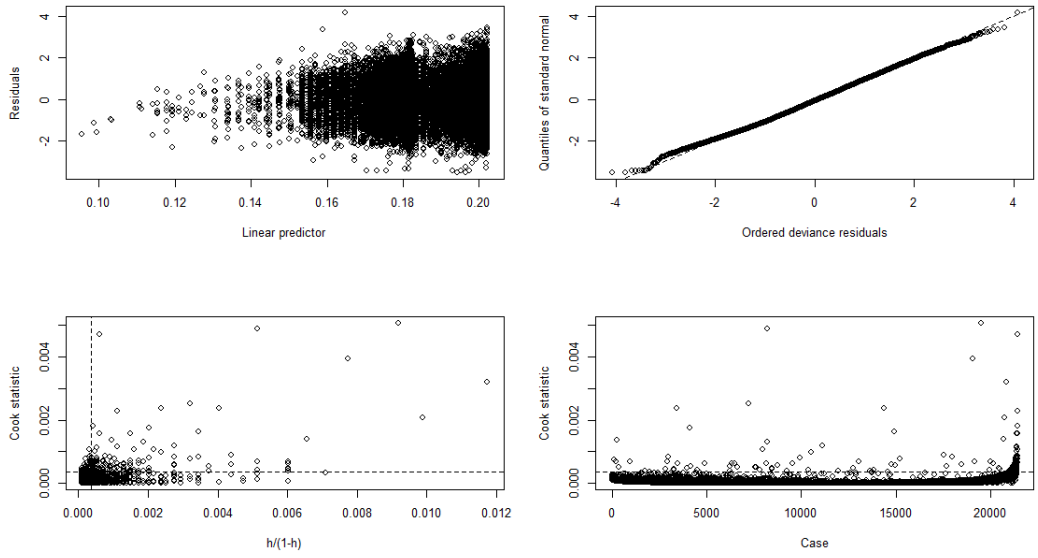
$$\text{Model 2: } s_1, x_2, x_2^2$$
$$\text{Model 3: } x_2, x_2^2, s_1 x_2.$$

Model 2 had an AIC of 65570.15 and model 3 had an AIC of 65706.30. It should be noted that at this point, our candidate models satisfy our intuition. Next, we ran diagnostics on these models as well as assessed their predictive power.

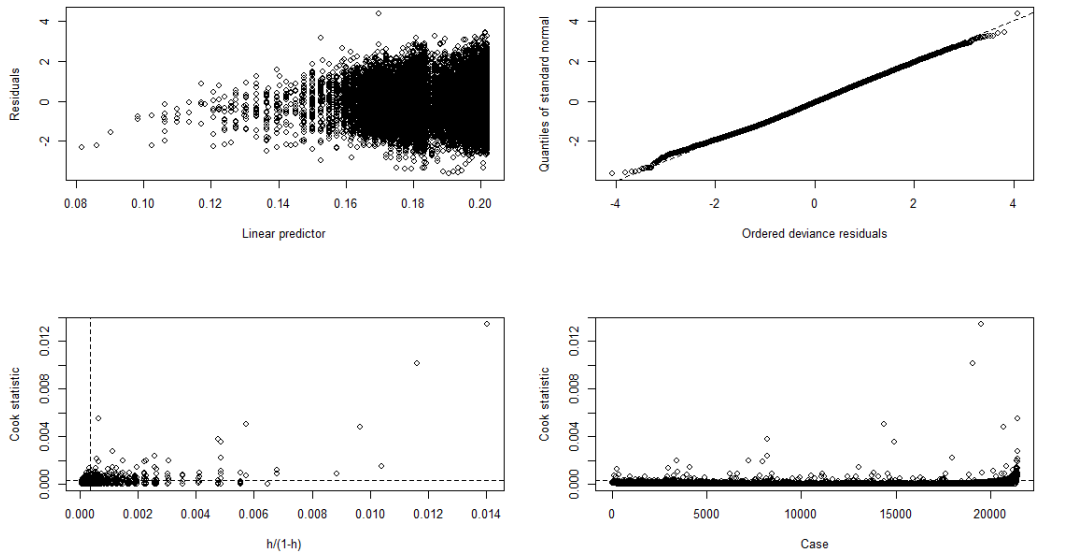
The diagnostics plot for model 1 is shown below:



The diagnostics plot for model 2 is shown below:



The diagnostics plot for model 3 is shown below:



There does not appear to be much difference between any of the three diagnostics plots. The main thing to note here is that high amount of observations that appear to be influential and high leverage for all three models. To help choose a model, we assess the predictive power of all three models.

We randomly chose 100 observations. We cross-validated multiple times to ensure the random sample of observations did not have much effect. In doing this, we saw that all three models had very similar predictive power.

Since all three models had similar predictive power, I chose model 2 as my final model for parsimonious reasons. It would be one of the easier models to explain since it does not have the interaction term. Although it had slightly higher AIC than model 1, it is one of the smaller models. This model can be specified as

$$\log\left(\widehat{\mu}_i\right) = 0.1392 - 0.01978 s_{i1} + 0.003382 x_{i2} - 4.54 \cdot 10^{-5} x_{i2}^2.$$

A shortcoming of this model would be the quadratic term of age, which would be difficult to explain to someone. Another shortcoming would be the negative coefficient on s_1. Intuitively, we would expect men to finish faster than women and thus if $s_1 = 1$, then we would expect to add time, not subtract. However, we met our intuition of both sex and age being related to finishing time.