STA 363 Project 1

Cassandra Hung

2024-11-01

Section 1: Linear Regression

One method that can be used to predict Parent-Child Internet Addiction Test (PCIAT) score is linear regression. In this section, we will build a linear regression model that uses all available features to predict PCIAT score. The features we have information on in this data set are as follows:

- `Age`: The age of the child in years.
- `Sex`: An indicator variable that is 0 when the child is male and 1 when the child is female.
- `SleepDistScore`: A child's sleep disturbance score, which is a numeric measure of how well the child sleeps. Higher scores indicate poorer sleep.
- `InternetUse`: A measure of how many hours a day the child spends on the Internet, with higher values corresponding to more hours per day.

This means that the model will include four explanatory variables: `Age`, `Sex`, `SleepDistScore`, and `InternetUse`. The goal of linear regression will be to create an equation that can be used to predict PCIAT score. Table 1 shows the resulting coefficients after building this model, and the corresponding least-squares linear regression (LSLR) line is shown below the table as Equation 1.

Table 1: Coefficients for LSLR Line

	Coefficient
(Intercept)	-6.84
Age	1.61
Sex	-4.65
SleepDistScore	0.34
InternetUse	3.53

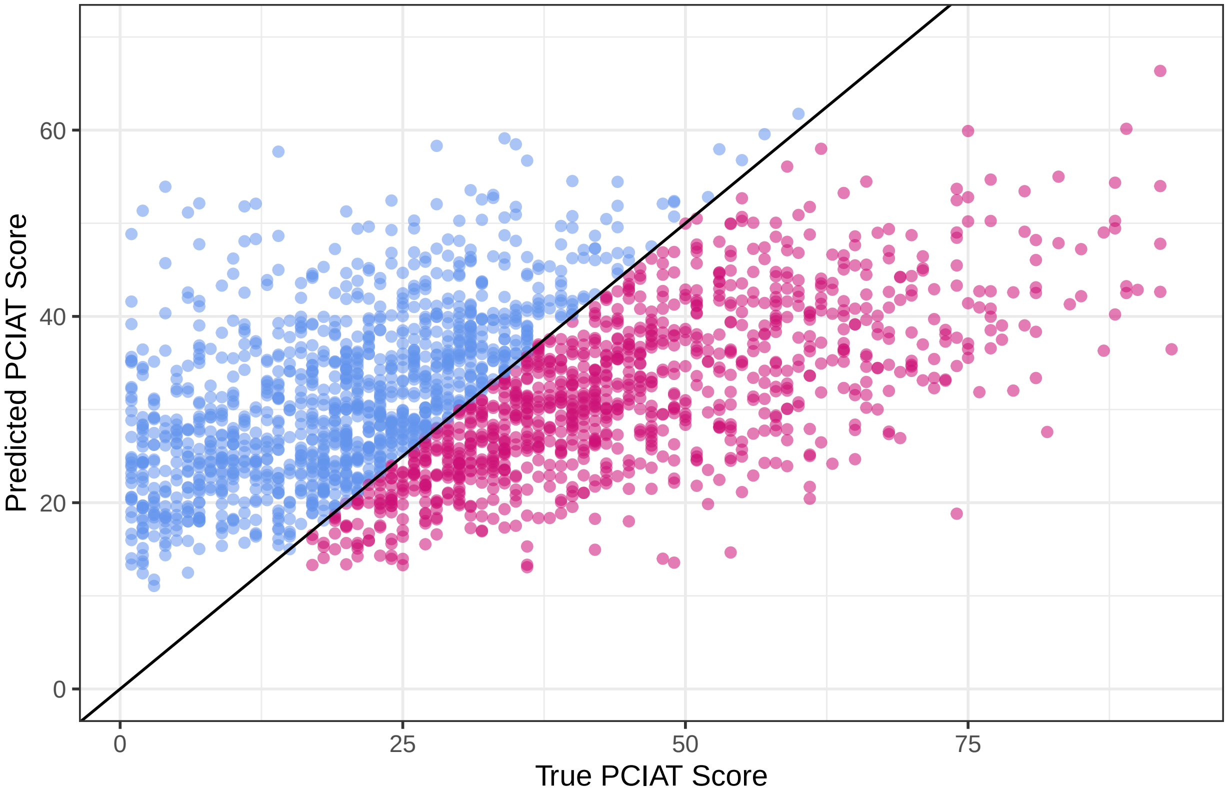
$$\text{Equation 1: } \widehat{\text{PCIAT}} = -6.84 + 1.61\text{Age} - 4.65\text{Sex} + 0.34\text{SleepDistScore} + 3.53\text{InternetUse}$$

The coefficients for the trained model shown in Table 1 provide some insight into how age, sex, sleep disturbance score, and internet use are associated with PCIAT score. For example, on average, it appears that increases in age, sleep disturbance score, and internet use are all associated with a higher PCIAT score, as evidenced by the positive coefficients associated with these features. This means that a child being older, having poorer sleep, and being on the internet more often are all associated with an increase in PCIAT score. However, the coefficient on sleep disturbance score is relatively small, meaning that on average, we expect only a small change in PCIAT score for every one-point difference in sleep disturbance score. Meanwhile, a child being female seems to be associated with a lower PCIAT score on average. Based on the magnitude of the coefficients, sex and internet use appear to have the greatest impact on PCIAT score. This is because on average, for every one-point increase in internet use, the PCIAT score is expected to increase by 3.53 points. Furthermore, PCIAT scores for male children are 4.65 points greater on average than those for female children. Overall, this model indicates that children who are older, male, have poorer sleep, and spend more time on the internet tend to have the highest PCIAT scores on average.

Now that we have trained a linear regression model, we will use leave-one-out cross validation (LOOCV) to assess its predictive accuracy. LOOCV is a cross validation technique that allows us to use one row in the data as validation data while using every other row as training data. We then repeat this process for every row in the data set so that each row gets to act as validation data once. While validation/train split is another cross validation method that is computationally simpler and faster than LOOCV, it is not appropriate for this data because this data set is relatively small. Splitting the data into validation and training sets will reduce the amount of data available to train the model. As a result, the predictive accuracy of the model may vary significantly depending on which data is randomly selected to be training data. Therefore, because we want to obtain a consistent estimate of the predictive accuracy of the model, we will use LOOCV instead of validation/train split.

To measure predictive accuracy, we will use root-mean-square error (RMSE), which is a measure of how far away our model's predictions are from the actual values in the data. After using LOOCV and linear regression, the RMSE is 16.33 points. This means that on average, our predicted PCIAT scores differ from the actual PCIAT scores by ±16.33 points. A scatter plot of predicted PCIAT scores versus true PCIAT scores is shown below in Figure 1.1, with predictions that are greater than the actual value shown in blue and predictions that are less than the actual value shown in pink.

Figure 1.1: Linear Regression Predicted vs. True PCIAT Scores



The plot in Figure 1.1 can be used to assess the quality of our predictions. In particular, we are interested in looking at where the data points fall relative to the black line shown in Figure 1.1. If a prediction and the true value were exactly the same, then the point would fall on this line. Thus, if the model could perfectly predict PCIAT score, then every point would fall exactly on this line. However, many points are spread out around this line, which means that there are many predictions that are not particularly close to the true score that we are trying to predict. In addition, the model tends to overestimate PCIAT score when the true PCIAT score is small, which can be seen in Figure 1.1 since most of the points representing a true PCIAT score below 15 fall above the line. Likewise, the model tends to underestimate PCIAT score when the true PCIAT score is large, as most of the points representing a true PCIAT score above 50 fall below the line. The model does provide better predictions for PCIAT scores between 15 and 50, as there are a number of points very close to the black line in the middle of Figure 1.1. Finally, the model almost always predicts a PCIAT score between 10 and 60. However, the actual PCIAT scores observed range from 1 to 93, so the true range and variability of PCIAT score is greater than what is being captured by our linear regression model.

Section 2: KNN

In addition to linear regression, K-nearest neighbors (KNN) is an algorithm that can be used to predict PCIAT score. In this section, we will explore using KNN for this purpose. To use KNN, we first assume that rows with similar values for their features will also be similar in their response variable. The algorithm first looks at Row 1 of the data and calculates the distance between Row 1 and the rest of the rows in the data. In this case, we will use Euclidean distance and all four features (age, sex, sleep disturbance score, and internet use) to calculate the distance between rows. Then, we pick the K rows (where K is a number that we choose) that are closest, or most similar, to Row 1 based on which ones have the smallest Euclidean distance to Row 1. Finally, we calculate the average PCIAT score for the K closest rows and use this average as the predicted PCIAT score for Row 1. This process is then repeated for all the other rows in the data.

One choice we have to make when using KNN is what value of K we will use. To do this, we will first consider all choices of K between 3 and 51. For each choice of K and using all the features, we will use LOOCV to assess the predictive accuracy of KNN with Euclidean distance to determine which value of K should be used. The results of this process are displayed in Table 2.

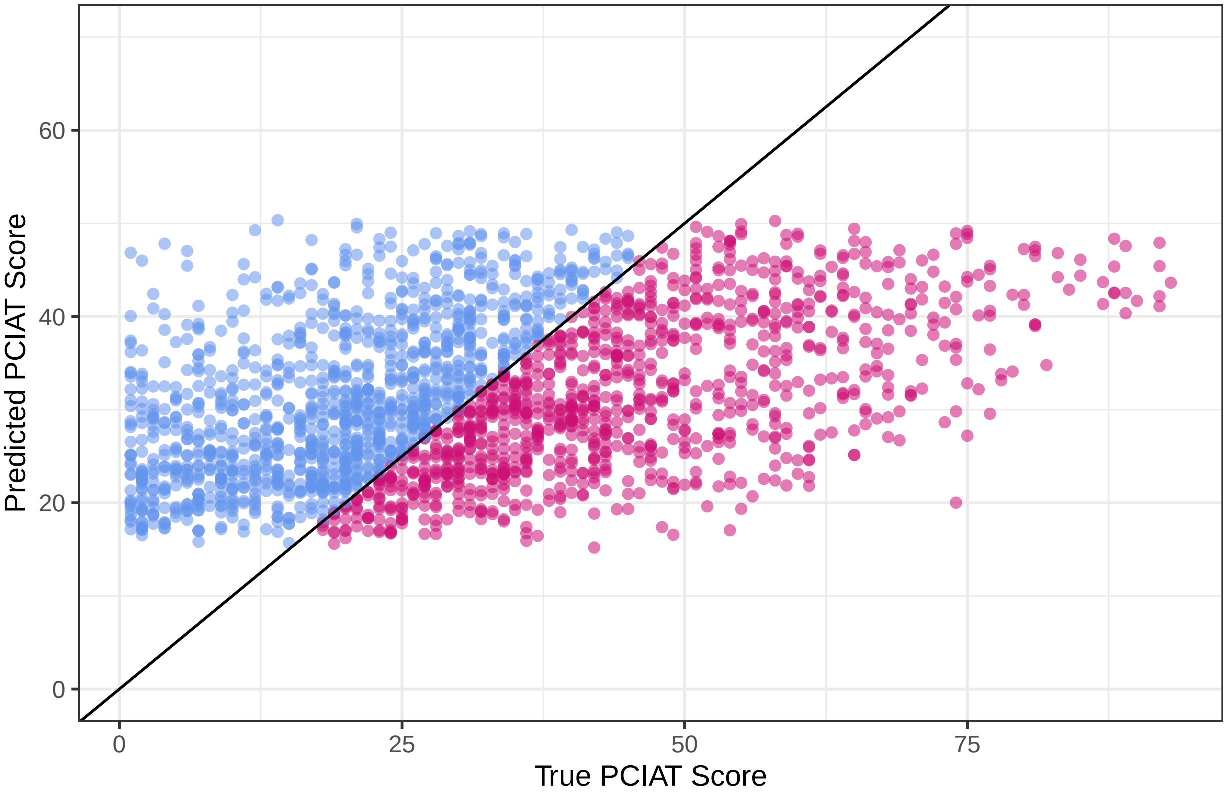
Table 2: Validation RMSEs for KNN

K	RMSE	K	RMSE
3	17.66	28.00	16.65
4	17.56	29.00	16.62
5	17.35	30.00	16.61
6	17.23	31.00	16.63
7	17.09	32.00	16.64
8	17.02	33.00	16.64
9	16.94	34.00	16.64
10	16.90	35.00	16.63
11	16.89	36.00	16.62
12	16.82	37.00	16.62
13	16.78	38.00	16.62
14	16.77	39.00	16.63
15	16.78	40.00	16.62
16	16.74	41.00	16.61
17	16.73	42.00	16.61
18	16.69	43.00	16.62
19	16.69	44.00	16.61
20	16.68	45.00	16.61
21	16.66	46.00	16.60
22	16.66	47.00	16.60
23	16.64	48.00	16.60
24	16.64	49.00	16.59
25	16.66	50.00	16.58
26	16.67	51.00	16.57
27	16.66		

The smallest validation RMSE in Table 2 is 16.57 points, which is the validation RMSE associated with K = 51. This means that on average, our predicted values for the PCIAT score differ from the actual scores by ±16.57 points when using 51-nearest neighbors. Another common choice of K to consider is $\sqrt{n}$, where $n = 2180$ for this data set. In this case, $\sqrt{n}$ is approximately 47. As such, we will choose between $K = 47$ and $K = 51$. In general, our goal is to minimize RMSE, but we should also consider what increasing K would mean for how efficiently we compute predictions. This is because using larger values of K means that more information is being used to calculate the average PCIAT score in the last step of the KNN algorithm, so using a smaller value for K when possible is more efficient for computations. However, we will prioritize the quality of our predictions in this case, as there is only a difference of four additional terms being used to calculate the average when using $K = 51$ as opposed to $K = 47$. Thus, KNN using $K = 51$ still performs better than the "default" choice of $K = 47$ since it is associated with a smaller validation RMSE. As such, $K = 51$ with an RMSE of 16.57 points is a reasonable choice for this data and is what we will use going forward.

Finally, we will visually assess our predicted PCIAT scores using 51-nearest neighbors with a scatter plot, shown below in Figure 2.1.

Figure 2.1: 51−NN Predicted vs. True PCIAT Scores



Looking at Figure 2.1, we again see a fairly even number of over-predictions and under-predictions. More specifically, the model appears to overestimate PCIAT score when the true PCIAT score is small, which is demonstrated in Figure 2.1 since most of the points corresponding to a true PCIAT score below 20 fall above the black line. Likewise, points representing a true PCIAT score of 50 or greater are all underestimates since they fall below the line. The model provides better predictions when the true PCIAT score is between about 20 and 45. Similar to when linear regression was used in Section 1, there is also more variation in the true PCIAT scores compared to the predicted scores. This can be seen in Figure 2.1 since the data points are farther from the line for lower and higher values of the true PCIAT score. Furthermore, this effect is even more noticeable using KNN with $K = 51$, as all the predicted scores fall between 15 and 55. As such, KNN struggled with predicting particularly low and high values of the PCIAT score and performed better when the true PCIAT score was between about 20 and 45.

Section 3: Conclusion

At this point, we have explored two methods to predict PCIAT score: linear regression and K-nearest neighbors using $K = 51$. To assess predictive accuracy, we calculated validation RMSEs and created scatter plots of the predicted and actual PCIAT scores for each model. The linear regression model with age, sex, sleep disturbance score, and internet use has an RMSE of 16.33 points, meaning that on average, predicted PCIAT scores using the regression model differed from the actual values by ±16.33 points. Meanwhile, the RMSE for KNN with $K = 51$ has a validation RMSE of 16.57 points. This means that the predictions produced using linear regression are slightly closer to the true PCIAT scores on average than the predictions from KNN. Therefore, linear regression should be used for the best predictive accuracy because the RMSE associated with linear regression is smaller than that for KNN.

However, while linear regression is the choice that yields better predictive accuracy when predicting PCIAT scores using this data, it is important to consider the quality of these predictions. Firstly, predictions that differ from the actual values by 16.33 points on average still represent a large difference between a prediction and an actual score, especially since we only observe PCIAT scores from 1 to 93 in this data set. Additionally, since RMSE describes the differences between predicted and actual values on average, it is the case that many predictions differ from the true score by more than 16.33 points. This can be seen in the scatter plot in Figure 1.1, where many of the points do not fall near the line. Another concern with these predictions is that they have a much smaller range than the true values, suggesting that the model does not provide high-quality predictions for lower and higher values of PCIAT score. In general, the linear regression model fails to capture the true variability in PCIAT score, and it could not predict PCIAT scores below 10 or above 70. Looking at Figure 2.1, KNN also has this issue. In fact, the problem is even further emphasized when using KNN as the predicted values have an even smaller range than the predictions using linear regression. As such, linear regression also performed better than KNN in terms of the range of PCIAT scores that could be predicted, but a reasonable next step would be to see if there are any ways to further improve the quality of these predictions. Methods that better capture the full range and variability of PCIAT scores would be particularly helpful, as that is one of the main concerns when looking at the current results using linear regression to predict PCIAT score.