Mini Project: Part III Samantha Maticka

0 Recap:

My project is looking at the relationship between covariates that describe different physical activities, namely swimming, biking, running, and walking. The covariates consist of variables detailing speed, distance, elevation change, etc. throughout the respective activity. Each activity is a single observation. There were ~700 runs, ~600 bikes, ~70 swims, and ~30 walks total.

For the linear regression the response variable that I am trying to predict is the start time of each activity in units of decimal-hour-of-day. For the logistic regression I am trying to predict the correct activity (i.e. 1 of 4 classes).

To see a list of all the original covariates see Table A.1 in the Appendix.

1 Prediction on the Test Set

Estimates of prediction error, MSE and accuracy, were both calculated by performing cross-validations across 10 folds of the training data. Test error and estimates of prediction error are listed in Table 1.

*Model selection method:* For the linear regression, a backwards BIC method was applied to all individual and 2-way interaction terms to minimize MSE whilst also minimizing complexity. For logistic regression the R-package ‘glmnet’ was used. The model-selection function, ‘cv.glmnet’ determines the ‘best’ model through an iterative process. Different lambdas are used to penalize larger coefficients in a maximum likelihood estimate. The model corresponding to a lambda that yields the lowest mean CV error was chosen. I used a ridge regression method for lambda, and the iteration was conducted 100 times (enough for error convergence). Chosen models are in Table A.2 of Appendix.

A/B: Take the best model you built for regression and for classification, and apply it to the test set you held out in the first part of the mini-project. Note the test error in each case; how does it compare to the estimate for test error that you derived previously (in part 2-c of your previous assignment)?

We had asked you to think about whether your process might have made you optimistic in your estimate of test error. One way to see this is to compare the test error to the *first* models that you built (without the benefit of having compared many models on the training data already).

With that in mind, compute the test error of the linear and logistic regression models that you built in part 1-c of the last assignment on your held out test set. How well do your previous estimates of test error for these models compare to what you find on the test set?

*Linear Regression:*

A/B: The best model for prediction of the continuous response variable (start time of activity) yielded **RMSE = 3.38 hours**. This suggests that on average the model predicts within 4 hours of the real start time. This isn’t very good, given that the athlete that this model was trained on would only start any given activity within about a 15 hour range (~ 5 AM to 8 PM).

Nonetheless, the model is used on the test data to get a true prediction error. **RMSE = 3.57 h**. This isn’t much different from the slightly optimistic estimate of prediction error given by the chosen model. The test error is comparable to the basic model design with a few covariates (1c) (**RMSE = 3.55 h**), and only slightly better than the ‘dumb’ model of predicting the mean (**RMSE = 3.75 h**).

*Logistic Regression:*

A/B: Since, we care about the accuracy of prediction per activity (e.g. TPrun/(total # of runs)), and not just overall model accuracy, the metric used to measure the model’s prediction ability was the average accuracy of the 4 individual accuracies. The chosen model does estimate an optimistic average accuracy (**98%**) of the test accuracy (**83%**), but the 2 base models are both well below (**50%** and **25%**) the test accuracy. The test accuracy of the base models compares to the estimated test error of each model respectively.

*Summary*: The overall prediction capacity of the linear regression model is nothing to write home about; the RMSE for all model cases (3-4 h) is on the order of the range in which the data exists (24 h). This is likely due to the nature of the response variable (start time of activity), which may inherently exhibit large variance. Admittedly, this dataset was chosen with the classification problem in mind – start time of activity seemed to be the best candidate for a continuous response variable. The prediction ability of the classifier does quite well, however, it is much better at predicting ‘runs’ and ‘rides’ than ‘walks’ and ‘swims’. This is likely due to the smaller availability of the latter 2 activities, as well as the similarity of characteristics between them (e.g. both are slow speeds, and may both be on flat land, etc.).

Table 1: Prediction Error from Test Data and Test Error.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model Name | Estimate of  Prediction Error  (MSE, h2 / Accuracy, %) | Test  Error  (MSE, h2/ Accuracy, %) |
| Linear Regression | Base 1 - mean | 14.10 | 14.63 |
| Base 2 – few covar | 12.64 | 12.81 |
| Chosen model | 11.45 | **12.77** |
| Logistic Regression | Base 1 - mean | Total: 51.58 %  Walk: 0 %  Swim: 0 %  Run: 100 %  Ride: 0 %  **Average Accuracy: 25 %** | Total: 50.17 %  Walk: 0 %  Swim: 0 %  Run: 100 %  Ride: 0 %  **Average Accuracy: 25 %** |
| Base 2 – few covar | Total: 92.66 %  Walk: 0 %  Swim: 5.26 %  Run: 99.83 %  Ride: 96.73 %  **Average Accuracy: 50.46 %** | Total: 91.81 %  Walk: 0 %  Swim: 0 %  Run: 100 %  Ride: 96.06 %  **Average Accuracy: 50.46 %** |
| Chosen model | Total: 98.21 %  Walk: 80 %  Swim: 91.23 %  Run: 99.01 %  Ride: 98.57 %  **Average Accuracy: 98.57 %** | Total: 96.25%  Walk: 37.50 %  Swim: 100 %  Run: 98.64%  Ride: 96.85 %  **Average Accuracy: 83.24 %** |

2 Inference

In this part of the project, you will try out some of the methods we learned in the class for *inference*. You should pick ONE of either the linear or logistic regression models that you built in part 2-a of the last assignment.

a)  For your chosen model, look at which coefficients are significant in the regression output, according to R. In words, describe what statistical significance means for these coefficients. Do you believe the results? Why or why not?

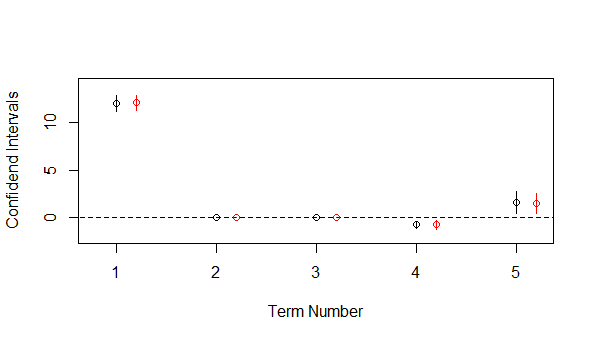
A: The Base 2 linear regression model will be analyzed. This model used 4 covariates. R gives a coefficient for each, and whether it is ‘statistically significant’. For each coefficient, R tests the null hypothesis that it is equal to 0. Larger p-values indicate a higher probability of accepting the null, and thus deeming the coefficient not significantly different from 0. The coefficients that are ‘significant’ are labeled with the confidence interval that they fall into (i.e. if this experiment were conduct many times, the null would be rejected *x* % of the time. Where, *x* is 99.9, 99, 95, 90, or 0). My caution with believing this is that the data I’m using has clustering in it and is thus, not strictly normal. Normality is assumed for the OLS procedure in R.

B: Now fit your chosen model on the test data, and look at whether the same coefficients are significant. Did anything change in doing so? If so, explain any differences that you found, and reflect on why they might be there.

When the model is fit to the test data, 2 of the covariates that were 99.9% significant with the training data became insignificant (not statistically different from 0). This sensitivity could indicate that these 2 covariates depend on other terms (not included) that the response variable *is* correlated to, or plainly, that the covariates used aren’t part of the true population model.

C: Use the bootstrap to estimate confidence intervals for each of your regression coefficients. Do the results differ from what R gave you in its standard regression output? Again, explain any differences that you found, and reflect on why they might be there.

The 95% confidence interval results from bootstrapping are remarkably similar to those reported by R (see plot below). They aren’t identical (so not likely a coding error). Both R and bootstrapping suggest Terms 2 and 3 are statistically insignificant. This is seen by their confidence interval encompassing a value of zero. If we were to plot 90% confidence intervals, terms 4 and 5 would capture 0 in their intervals. Two methods with the same result is reassuring of the result.



Term Number

Bootstrapped: B = 100

R output

D: Compare the significant coefficients in the model from 2-a of the last assignment, to the significant coefficients in the model from 1-c of the last assignment (your first regression model). Did the significant coefficients change? If so, explain the differences.

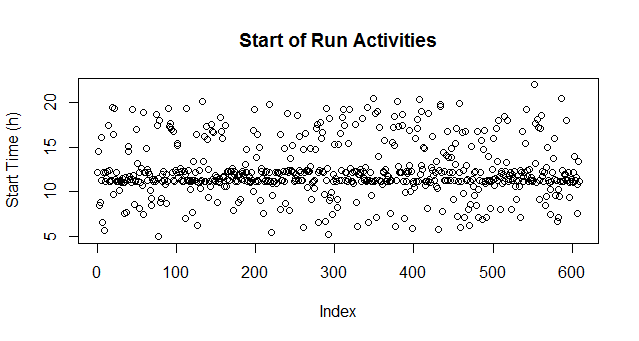
Realizing I used the wrong model for part C (I used part II-1c), I’ll now compare it to the same coefficients in part II-2a. In part 1c (4-covariate model) the significant terms are ‘*total time of activity’* and a measure of ‘turns/curves’, in part 2a these are no longer significant - only interaction terms show significance. Specifically, *Umed:Dispmed, Umed:dEmax, Dispmed:Turnsmed, Dispmed:Grademed, dEmax:Grademed* (descriptions in Table A.1 of Appendix). Neither of the 2 significant coefficients in part 1c seems to physically depend on the interaction terms, so collinearity doesn’t seem to explain it. It’s likely that the response variable chosen for this project was just a poor choice and the covariates we are using aren’t in the true population model.

E: Comment on potential problems with your analysis, including collinearity, multiple hypoth- esis testing, or post-selection inference; be specific. For example: what evidence suggests to you that collinearity is impacting your results? For multiple hypothesis testing, suggest how applying the Bonferroni correction would change your interpretation of significant coeffi- cients. For post-selection inference, suggest aspects of your model-building and inference process that might bias your determination of which coefficients are significant.

Obvious collinearities (e.g. median vs mean speed) were reduced to 1 covariate before training. There is a possibility that collinearity exists among the remaining covariates, since all were created from latitude, longitude, elevation, and timestamps. However, since mathematical manipulations were used to create covariates, it isn’t obvious which ones may be collinear – at least none are scaled versions of each other. For linear regression with backwards BIC, the subset of covariates that yields the lowest MSE was chosen. Since this is conducted with cross-validation, it minimizes the effect of favorable selection since, presumably the covariates minimized MSE in all of the k-folds. This suggests that covariate selection *may* be extrapolated to the broader population (if the training data is a good representation of the population). For logistic regression, ridge regression was used, and the lambda that yielded the lowest cv MSE was chosen, but the subset of covariates was chosen by trial and error of adding/removing covariates to maximize average accuracy. It’s likely that this favorable selection led to the optimistic prediction error estimate and possible removal of relevant covariates.

F: For the relationships that you found that are significant, would you be willing to interpret them as causal relationships? If not, why not? What other covariates do you think might be confounding your ability to infer causal relationships?

Of the relationships that are significant in the final model, some of them I would interpret with caution as causal relationships, and others I would think of as solely for prediction. For instance, a *‘run-activity indicator’* has a strong relationship to *‘start time of activity’*. This could be physically represented by the athlete having a running routine – I happen to know this athlete often runs on his lunch break. Given that the activity is a run, noon is a good guess for the start time. This is seen in the figure below:



Other variables, such as *‘median speed’* I wouldn’t necessarily expect to dictate the *‘start time’* or vice versa, since the data is described by 4 separate activities – each of quite different speed ranges. It’s possible that start-time-preferential is dictated by activity type, and if activity type dictates the activity speeds, then it’s a collinearity that would describe the relationship.

3 Inference

Finally, we would like you to summarize your work on the project. This is open-ended, and so you should choose what you think are the most interesting or important things to report. The following high-level questions are meant to guide your thinking.

a) How would you expect your models to be used practically? Do you think they would primarily be used for prediction, for inference, or for both? What decisions do you think your models would guide, and what pitfalls do you see in using your models to make these decisions?

b) How well would your models hold up over time (i.e., how often do you think they should be refitted)? Why?

c)  Are there choices you made in your data analysis, that you would want to make sure any one (e.g., a manager, a client, etc.) that uses your models is aware of? Examples here might include approaches to data cleaning; data transformations that you chose; vulnerability to overfitting, multiple hypothesis testing, or post-selection inference; etc.

d)  If you could, how would you change the data collection process? In particular, are there reasonable covariates you would like to collect, that were not present in the data?

e)  If you were to attack the same dataset again, what would you do differently?

A: The *linear regression* model in this project was designed more with inference in mind – out of curiosity (and project requirement to do a linear regression) it would be neat to know what characteristics of a workout might determine when the activity was initiated. This model has a serious pitfall, namely, its error is similar to the most basic model of predicting the mean. I believe this is due to the nature of the chosen response variable – it depends on too many unpredictable things (e.g. athlete’s behavior/preference/un-steady routines.. ).

The *logistic regression* was designed only with prediction in mind. Overall, I think this model is a good starting point and a beacon of hope that it could evolve to something more robust, but it has its limitations. To summarize, I’ll break it into 1) the limitations I’ve discovered and 2) some potential extensions that could improve the predictive power.

Limitations:

1. *Sampling bias per athlete*: this model was only trained on 1 athlete. So, while there is clustering per activity type, if more athletes are considered there is a second dimension of clustering related to athletic ability.
2. *Sampling bias per location:* While this is a well-traveled triathlete, most of his activities reside in the Bay Area, where there are ample hills. Take an athlete from Florida, and elevation won’t have the same ability to predict.
3. *Sampling bias activity:* There are far more runs and rides that walks and bikes. More of the latter are need for better tuning.

Extension:

1. Guide a predictive model that takes user-defined input. For instance, when an athletes starts using Strava the model will make a guess and based on certainty, it could ask the user ‘was this a swim?’ This would require training *per athlete*. It may be wise to re-train occasionally over the lifetime of app-usage in case an athlete ramps-up or drops-off their fitness. Not too frequent to annoy users with questions.
2. Geo Mapping the routes could help immensely with predictive power, by associated water and roads to swims and bikes, respectively.

If I were to attack the same dataset again, I would 1) give myself more time (R took more time than I had anticipated), 2) employed a naïve Bayes Classifier for continuous variables (assigning a normal distribution to continuous covariates, rather than empirical frequencies) mostly to have a behind the scenes look at the regression process, and 3) extend the dataset by collecting data from more friends.

*I really enjoyed the class. Thanks, Ramesh (and TAs)!*

Table A.1: Initial covariates to work with. Collinear covariates were include to determine the better metrics.

|  |  |  |
| --- | --- | --- |
| **Covariate** | **Description** | **Calculation** |
| Continuous | | |
| Umean | Average Speed | mean(speed) |
| Umed | Median Speed | Median(speed) |
| Umax | Maximum Speed | Max(speed) |
| Dtotal | Total Distance Covered | Sum(Distances) |
| dEmax | Maximum Elevation difference | Max(Elevation) – min(Elevation) |
| Grademed | Median Grade [path slope] | Median( |(ΔElevation)/Distance| )\* |
| Ttotal | Total moving-time of activity | Sum(Δt)\*\* |
| Turnsmed | Curviness/small proximity of route  (calculated per 10-min bins) - median | Median( max displacement within 10 min / total distance within 10 min) |
| Turnsmin | Curviness/small proximity of route  (calculated per 10-min bins) - minimum | Min( max displacement within 10 min / total distance within 10 min) |
| Turnsmean | Curviness/small proximity of route  (calculated per 10-min bins) - mean | Mean( max displacement within 10 min / total distance within 10 min) |
| Dispmed | Max displacement within a 10-min window - median | Max(displacement within 10 min) |
| Categorical | | |
| Activity | Type of activity (run, swim, bike, walk) | Known prior to processing |
| Iswim | Swim Indicator | Logical: (Umed < Swim Threshold)\*\*\* |
| Ibike | Bike Indicator | Logical: (Umed > Run Threshold)\*\*\*\* |

\*this is irrespective to whether it was an ascent or descent.

\*\*time steps that indicated no movement (speed < 0.3 m/s) were removed before processing. Therefor the sum of Δt is used as opposed to the difference between end and start time, to only account for moving time.

\*\*\*The median speed is used to account for outliers (e.g. if someone forgot to turn off there app after swimming and before biking home). Threshold choice: Michael Phelps swam ~1.9 m/s for 200 meters - assume if the median speed of activity is greater than this, the athlete is probably not swimming. Note: people can also walk this slow.

\*\*\*\*Usain Bolt ran 12.5 m/s for 100 meters – assume if the median speed of activity is greater than this, the athlete is probably not running, and thus, biking. Note: this is still fast for a biking pace, so may be too strict.

Table A.2

|  |  |
| --- | --- |
|  | Model |
| Linear Regression |  |
|  |
|  |
| Logistic Regression |  |
|  |
| + *Grademed:. + Iswim:. + Umed:.* |