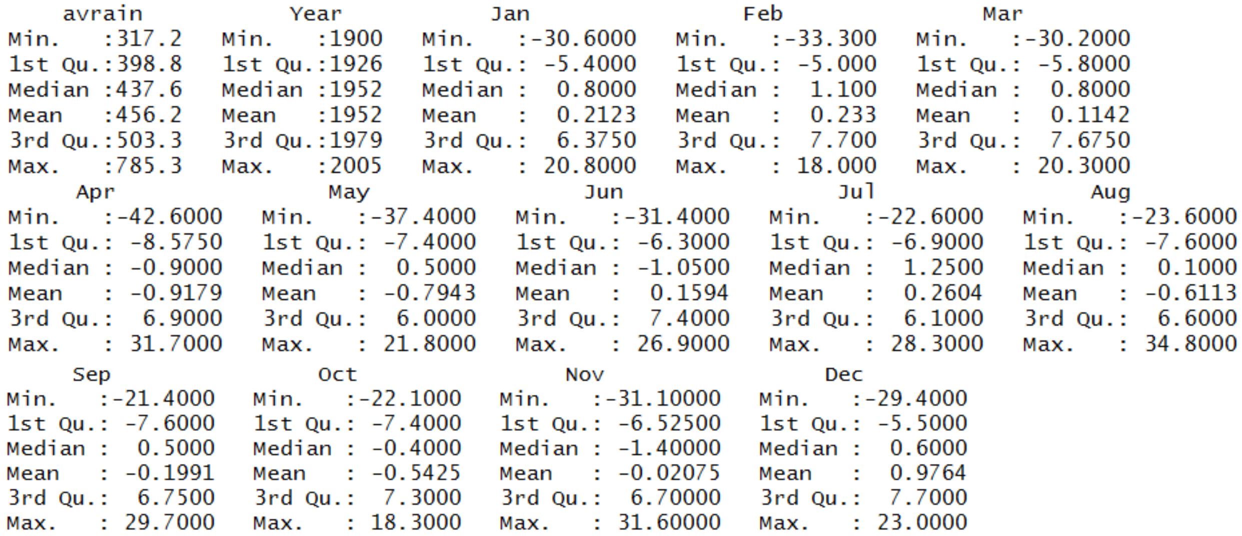
Making it Rain: Predicting the Weather in Australia
Arjun Gupta
4/29/15
Professor Richard Berk, STAT 474

We seek to gain some knowledge about the relationship between the average Southern Oscillation Index (SOI) in a given month, the year, and the average rainfall in Australia that year, as well to see how well we can forecast average rainfall in a given year. As rainfall is important for agriculture which affects the entire Australian economy and its populace, policy makers are likely interested in these questions. It is important to first note that climate science is complicated; it is impossible to expect that we have all the predictors necessary to fully explain it, and as statisticians with little subject-matter knowledge it is unlikely we would know the proper way to model it. Thus, we are necessarily in the wrong model perspective, where instead of trying to determine parameters of some true response function, we are instead estimating a population approximation to see what things we can learn, as well as discern associative relationships (rather than somehow claim we can understand causal ones).

Summary Statistics

With the wrong model perspective in mind, we begin with a preliminary analysis of the data:



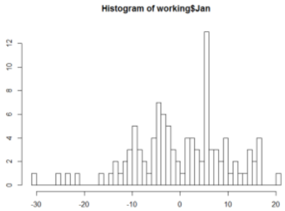
```
     avrain          Year          Jan               Feb               Mar
 Min.   :317.2   Min.   :1900   Min.   :-30.6000   Min.   :-33.300   Min.   :-30.2000
 1st Qu.:398.8   1st Qu.:1926   1st Qu.: -5.4000   1st Qu.: -5.000   1st Qu.: -5.8000
 Median :437.6   Median :1952   Median :  0.8000   Median :  1.100   Median :  0.8000
 Mean   :456.2   Mean   :1952   Mean   :  0.2123   Mean   :  0.233   Mean   :  0.1142
 3rd Qu.:503.3   3rd Qu.:1979   3rd Qu.:  6.3750   3rd Qu.:  7.700   3rd Qu.:  7.6750
 Max.   :785.3   Max.   :2005   Max.   : 20.8000   Max.   : 18.000   Max.   : 20.3000
      Apr                May                Jun                Jul                Aug
 Min.   :-42.6000   Min.   :-37.4000   Min.   :-31.4000   Min.   :-22.6000   Min.   :-23.6000
 1st Qu.: -8.5750   1st Qu.: -7.4000   1st Qu.: -6.3000   1st Qu.: -6.9000   1st Qu.: -7.6000
 Median : -0.9000   Median :  0.5000   Median : -1.0500   Median :  1.2500   Median :  0.1000
 Mean   : -0.9179   Mean   : -0.7943   Mean   :  0.1594   Mean   :  0.2604   Mean   : -0.6113
 3rd Qu.:  6.9000   3rd Qu.:  6.0000   3rd Qu.:  7.4000   3rd Qu.:  6.1000   3rd Qu.:  6.6000
 Max.   : 31.7000   Max.   : 21.8000   Max.   : 26.9000   Max.   : 28.3000   Max.   : 34.8000
      Sep                Oct                Nov                 Dec
 Min.   :-21.4000   Min.   :-22.1000   Min.   :-31.10000   Min.   :-29.4000
 1st Qu.: -7.6000   1st Qu.: -7.4000   1st Qu.: -6.52500   1st Qu.: -5.5000
 Median :  0.5000   Median : -0.4000   Median : -1.40000   Median :  0.6000
 Mean   : -0.1991   Mean   : -0.5425   Mean   : -0.02075   Mean   :  0.9764
 3rd Qu.:  6.7500   3rd Qu.:  7.3000   3rd Qu.:  6.70000   3rd Qu.:  7.7000
 Max.   : 29.7000   Max.   : 18.3000   Max.   : 31.60000   Max.   : 23.0000
```

Figure 1

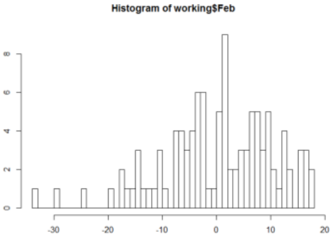
According to the government of Australia’s website (http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1301.0~2012~Main%20Features~Australia%27s%20climate~143) the units of rainfall are mm, and average annual rainfall in over 80% of the country is below 600mm and is less than 300mm in over 50% of the country. As our data are yearly “area-adjusted” averages, this is in line with our values, with the min and the max falling between 317.2 mm and 785 mm. As we can see, Australia is very dry. 50% of years had rainfall of 437 mm or less. To put this into perspective, that’s a median annual rainfall of just 17 inches.

Now we look at the five-number summaries of the Southern Oscillation Index (SOI) for each month. Nothing jumps out at first glance. The maximum SOI recorded was in August with 34.8, and a minimum in April of -42. While all months have outliers in the positive and negative directions, the 1st and 3rd values for each month fall between -8 and +8, suggesting SOI is generally pretty stable.

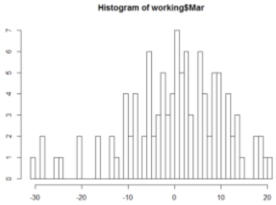
We also notice that the distribution of average rainfall is right-skewed (median of 437, mean of 456, and a 3rd quartile of 503) and that most of the predictors are left-skewed but not overly so: except for June, November, and December, each month's median SOI is higher than its mean. We will examine the histograms of each month to further look into this



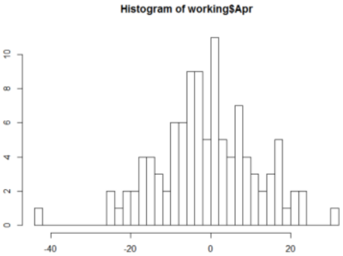
January



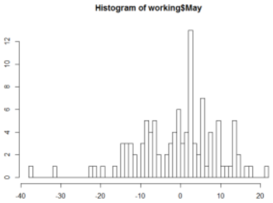
February



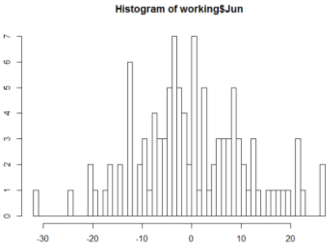
March



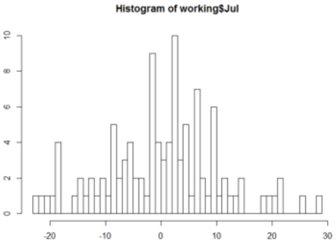
April



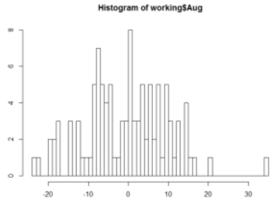
May



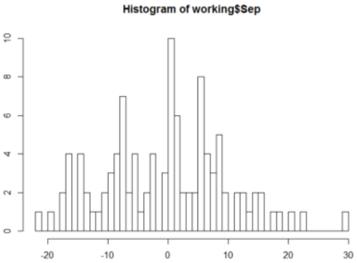
June



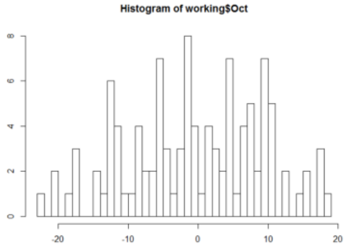
July



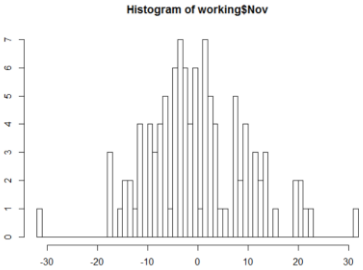
August



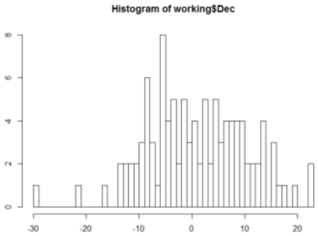
September

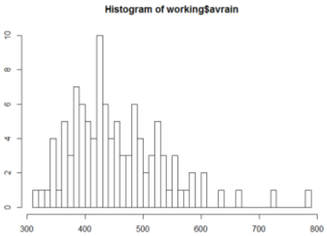


October



November





December Average Rain

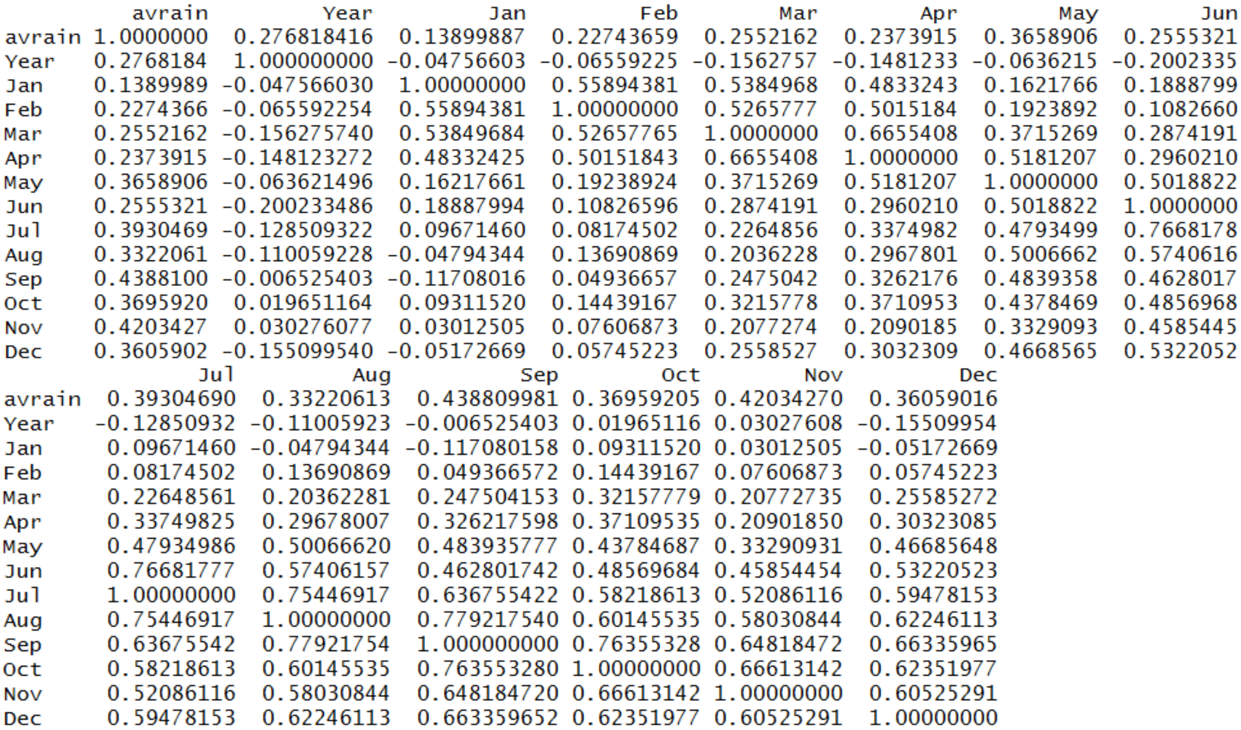
Figure 2

We examine the histograms of each month's average SOI values. The 5-number summaries told us that most of the distributions were somewhat left-skewed, which the histograms confirm. January through May all demonstrate right skew, and each have a couple very low SOI values. While June is actually slightly right skewed it does have an outlier at -31 millibars. In fact, almost every month has an value more extreme than ±20. Some months (November, July, and April) have outliers both on the left and right sides of the distribution. The distributions of September and October are interesting—there are clusters of values that occur with high frequency about every 5 or 10 millibars.

Do these values check out? Without deep subject matter knowledge it's hard for us to know; climate science is a complicated topic. But, skimming through these plots of the SOI since 1900, (provided by the government of Australia https://www.longpaddock.qld.gov.au/seasonalclimateoutlook/southernoscillationindex/soigraph/index.php) shows that the SOI seems to normally fluctuate between -10 and 10 millibars for any given month in a year, so each month having a few values spike to -42 or +34 in the past 100 years seems plausible. Again, we would consult climate scientists if we could, but within our means of internet research and intuition the data seem to check out without any serious quality flaws.

It is important to think about how what we learned from the histograms will affect our stochastic gradient boosting procedure. Using a base classifier of quantile regression trees yields one nice property: We are estimating conditional medians rather than means. What this means (no pun intended) is that outliers will not influence our results as strongly. Since the loss function is linear deviation rather than quadratic, the penalty on missing outliers is much less harsh, and our regression trees won't try as hard to bend to the whims of outliers when fitting. This makes our fits more stable, a plus when we only have 106 observations to work with. Recall that our regression trees are still splitting on ordinal data to reduce variance—as long as there are no miscodings, the non-normal distributions as well as outliers should not be an issue. These values far to the right and left of the median will simply fall to the right or left of whatever split our trees make. For the same reason, the fact that the response itself is only slightly right-skewed (the median is 437 mm, the mean 456) should not be a huge problem either. If the mean were significantly greater than the median, say by 200 mm, then we might be concerned, as regression trees do not perform as well on skewed distributions (there is less impurity to reduce) without cost-weighting. But for this dataset the skewedness and outliers don't concern us.

We also should check the correlations between our predictors—multicollinearity would certainly have an effect on our forthcoming analysis.



```
            avrain        Year         Jan         Feb        Mar        Apr        May        Jun
avrain 1.0000000  0.276818416  0.13899887  0.22743659  0.2552162  0.2373915  0.3658906  0.2555321
Year   0.2768184  1.000000000 -0.04756603 -0.06559225 -0.1562757 -0.1481233 -0.0636215 -0.2002335
Jan    0.1389989 -0.047566030  1.00000000  0.55894381  0.5384968  0.4833243  0.1621766  0.1888799
Feb    0.2274366 -0.065592254  0.55894381  1.00000000  0.5265777  0.5015184  0.1923892  0.1082660
Mar    0.2552162 -0.156275740  0.53849684  0.52657765  1.0000000  0.6655408  0.3715269  0.2874191
Apr    0.2373915 -0.148123272  0.48332425  0.50151843  0.6655408  1.0000000  0.5181207  0.2960210
May    0.3658906 -0.063621496  0.16217661  0.19238924  0.3715269  0.5181207  1.0000000  0.5018822
Jun    0.2555321 -0.200233486  0.18887994  0.10826596  0.2874191  0.2960210  0.5018822  1.0000000
Jul    0.3930469 -0.128509322  0.09671460  0.08174502  0.2264856  0.3374982  0.4793499  0.7668178
Aug    0.3322061 -0.110059228 -0.04794344  0.13690869  0.2036228  0.2967801  0.5006662  0.5740616
Sep    0.4388100 -0.006525403 -0.11708016  0.04936657  0.2475042  0.3262176  0.4839358  0.4628017
Oct    0.3695920  0.019651164  0.09311520  0.14439167  0.3215778  0.3710953  0.4378469  0.4856968
Nov    0.4203427  0.030276077  0.03012505  0.07606873  0.2077274  0.2090185  0.3329093  0.4585445
Dec    0.3605902 -0.155099540 -0.05172669  0.05745223  0.2558527  0.3032309  0.4668565  0.5322052
              Jul         Aug          Sep        Oct        Nov         Dec
avrain  0.39304690  0.33220613  0.438809981 0.36959205 0.42034270  0.36059016
Year   -0.12850932 -0.11005923 -0.006525403 0.01965116 0.03027608 -0.15509954
Jan     0.09671460 -0.04794344 -0.117080158 0.09311520 0.03012505 -0.05172669
Feb     0.08174502  0.13690869  0.049366572 0.14439167 0.07606873  0.05745223
Mar     0.22648561  0.20362281  0.247504153 0.32157779 0.20772735  0.25585272
Apr     0.33749825  0.29678007  0.326217598 0.37109535 0.20901850  0.30323085
May     0.47934986  0.50066620  0.483935777 0.43784687 0.33290931  0.46685648
Jun     0.76681777  0.57406157  0.462801742 0.48569684 0.45854454  0.53220523
Jul     1.00000000  0.75446917  0.636755422 0.58218613 0.52086116  0.59478153
Aug     0.75446917  1.00000000  0.779217540 0.60145535 0.58030844  0.62246113
Sep     0.63675542  0.77921754  1.000000000 0.76355328 0.64818472  0.66335965
Oct     0.58218613  0.60145535  0.763553280 1.00000000 0.66613142  0.62351977
Nov     0.52086116  0.58030844  0.648184720 0.66613142 1.00000000  0.60525291
Dec     0.59478153  0.62246113  0.663359652 0.62351977 0.60525291  1.00000000
```

Figure 3

It is worth remembering that since Australia is in the Southern hemisphere, their seasons are flipped:

- Spring - the three transition months September, October and November.
- Summer - the three hottest months December, January and February.
- Autumn - the transition months March, April and May.
- Winter - the three coldest months June, July and August.

(Quoted from http://www.bom.gov.au/climate/glossary/seasons.shtml)

As we expect, months of the same season are likely to be similar to each other—how different is July from August year after year? Indeed, the correlation between December and November is 0.6, the correlation between December and October is 0.623, the correlation between November and October is .66, June and July have a correlation of .76, etc.

What is slightly more surprising is the strength of correlations between months that are in different seasons. December is the beginning of summer, and July is the middle of winter, yet the correlation between the two months is a high .59.

There may be a couple reasons for this. The persistence of various climate trends that last for years probably linger in our data (which are set up not to be time series, but just average SOI for each month of each year). Additionally, it’s also possible that the extreme seasons (winter and summer) share traits that the transition seasons don’t, and vice versa. And of course, when a given year happens to be more mild overall (with less extreme temperatures, rainfall, SOI in every season), that would make the months of that entire year more correlated, adding to correlation.

Upon closer examination, there's an interesting pattern. We find that of all the pairwise correlations for the months May through December, none of them have a correlation less than 0.3, and most months that are not in the same season are 0.4 and above. May through December takes us through Australia's winter, spring, and 1/3 of the way through summer. Interesting. One would not expect December to have a correlation of 0.6 with July.

Then, January through April buck this trend. Within this range (January, February, March, April), all pairwise correlations between these months is never less than 0.4. Each of these months almost never has a correlation above .35 with the months mentioned in the paragraph above (the only exception being April having on .51 correlation with May, understandable as they are in the same season). January and February are either negligibly or negatively correlated with the other months mentioned, which for some is quite strange. Why do January and February have either 0 or negative correlation with December, when all three are summer months?

This handy plot, produced by the package "corrplot," sheds some answers. The darker blue, the higher correlation between any two months. We have relabeled January as x2, February as x3, and so on:

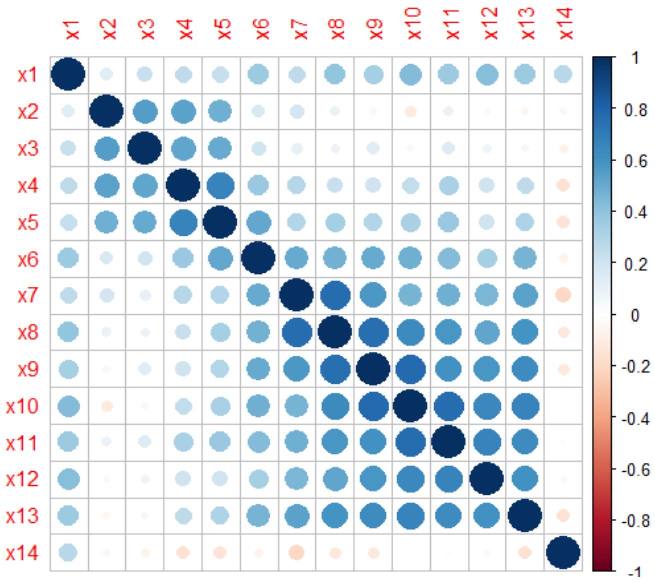


Figure 4

The El Niño/La Niña months are highly correlated with each other, and the non-El Niño/La Niña months are correlated with each other. It turns out that the El Niño/ La Niña phenomenon, when it occurs, reaches peak strength around January/December until April (http://oceanservice.noaa.gov/facts/ninonina.html)—exactly the period we see is correlated with each other, and not as correlated with the other months. El Niño/La Niña can sometimes last for 3-5 years, which also aligns with the idea that correlations between months overall will be high when trends persists over years.

Of course, these explanations don't capture everything. March, which is the beginning of autumn and still an El Niño month, has a correlation between .2 and .3 for all the other non-El Niño months, not exactly in accordance to our purported explanation. It's important to remember we there are likely much more complicated reasons and forces at play that cause our month-on-month correlation values. If we were doing this in real life, we would check with the subject-matter experts about whether these findings are reasonable. But based on what we can piece together, the data again checks out as valid with no serious flaws. And importantly, we have discovered that the majority of our predictors are pretty correlated.

We should think about how this collinearity will affect stochastic gradient boosting. Unlike random forests, stochastic gradient boosting considers all possible predictors at each iteration when fitting a regression tree, rather than randomly selecting them. What this means is that all of the predictors are competing with each other every time—those that are highly correlated to each other are similar likely account for the same portion of signal in the response. This will cause the one that is slightly better on average to "crowd out" the other, meaning the first predictor may be split on much more often, and the second (only slightly weaker predictor) will not get to participate. This means we could 1) miss highly local signal in the response that the second predictor might have helped us capture better, and 2) that our variable importance plots may come out strangely--it will appear that some months consistently don't matter when in fact they are simply highly correlated with another predictor. A potentially comforting thought is that if we are sampling residuals randomly, over a large number of iterations perhaps the highly local portions of the data will get selected, and those slightly weaker predictors will get to participate. But it is possible that we miss out on a small bit of signal, and have less interpretable variable importance plots because of the way stochastic gradient boosting chooses from all predictors at each iteration.

Year simply runs from 1900 to 2005, and as mentioned in class, the data are presented as to not make them a time-series. There is nothing to interpret in the year histogram.

With this in mind we see that while each predictor distribution has outliers, most of the predictor values have interquartile ranges of around -8 to +8. The distributions of the response and the predictors exhibit skew, but as noted about our regression trees with linear loss functions should not have issues with this. Predictors are highly collinear, something that may cause concern later on.

Implementing GBM

The GBM package was developed to allow us to implement Friedman's stochastic gradient boosting method. What this procedure does is initialize a predicted value for all observations (like the grand mean), take a random sample without replacement of the resulting residuals (defined by some loss function, in our case linear), run some sort of classifier on them, and use the fitted values to update the original fitted values. It then computes new residuals from the updated fitted values, randomly samples another portion of the data (in accordance with the idea that introducing randomness into the fitting process can counterintuitively improve performance, something first proposed by Breiman), and continues iterating in this fashion. The stopping criteria is determined by seeing how the algorithm is doing on our best approximation of test data at each iteration. For example, by using something like CV folds. CV folds split the data into k subsections, train a model on k-1 out of the k subsets, and see how the model performs on the held out subset. It does this k times (k choose k-1=k) and the takes an average of the performances. If the performance from iteration to iteration isn't improving enough, the algorithm stops and outputs fitted values.

GBM has a variety of tuning parameters, and using quantile regression trees as our base classifier we must also decide which quantile to estimate—which means determining a relative cost ratio. Thus, we have a variety of tuning parameters to specify. Let's get started.

With only 106 observations, it is probably not a good idea to split the data into training, evaluation, and test sets like we might normally do with a larger data set. Particularly after working with random forest, we very much like the idea of using out of bag data to see how much reduction in deviance we are getting on each iteration. However, in the GBM guide, Ridgeway explains that using reduction in deviance based on out of bag data is "almost always" too conservative in estimating the optimal number of iterations (8). Furthermore, in *Statistical Learning from a Regression*, Berk goes on to explain that since the *point "*is to determine how well the iterations are doing with the data on hand, it is not clear that a more conservative estimate is called for" (277). We see no reason not to heed Ridgeway's advice, and will use cross-validation folds to determine to determine the optimal number of iterations. With only 106 observations, we will really need to squeeze this data set for all the signal we can get out of it, meaning we probably prefer more iterations over less.

As recommended by both Berk (271) and Ridgeway (8), we will set bag size to the standard 0.5 of the data.

With bag size=0.5 on a dataset with 106 observations, 53 observations will be randomly sampled without replacement on each iteration. We will set the cv.fold argument to 3. Using 34 observations to train a tree and 19 observations to see how much deviance was reduced seems reasonable. With such a small data set, it doesn't seem like a good idea to set cv.fold to something like 10—5 observations would be too few to validate on meaningfully.

As for lambda, in class the default setting was .001. Having such a small lambda allows us to decrease bias, for we are taking steps along the gradient function very slowly. If we allow our fitted values to increase only very incrementally, we will protect against overfitting because each time we are only allowing the fit to be updated very slightly which each iteration of the data—we are muting the amount of idiosyncratic noise that we are taking advantage of from any given sample. We are able to compute a larger number of basis functions and the fitting function can be more flexible. In fact, Berk goes on to say that values of lambda smaller by a factor of 10 are worth a try as well (271). Using a relatively new computer, computational power should not be an issue for us. We set lambda=.0001--Ridgeway explains that the smaller the lambda one can afford computationally, the better (7).

Ridgeway suggests setting n.trees to anywhere between 3,000 and 10,000. Again, since computational power is not an issue, we will opt for 10,000.

Now we come to the two most tricky tuning parameters, interaction depth and min bucket size. The two relate to how large of trees we might build. All else constant, a low minimum bucket size should yield a larger tree, and a high interaction depth should also yield a larger tree. As Professor Berk said in class “bucket size is just a hack to use when you don’t know how deep interaction effects should be.” And we are inclined to agree. Particularly with such a small data set, what is an appropriate final bucket size? 5? 10? 15? It’s difficult for us to really reason out, and we’d rather let it go to the default minimum bucket size of 10.

However, reasoning with interaction depth is easier. Climate science is complicated, so using an interaction depth of only 1 is probably too simple to capture all the mechanisms going on. We might imagine that average rainfall is some function of month on month differences (perhaps a low SOI in December, a high SOI in January, all occurring in a certain year yields a different amount of rain than starting with high SOI in December)—this formulation would certainly require multiple levels of interaction effects. We don’t want to overcomplicate things and give our trees too many basis expansions to choose from, but it would be prudent to at least try and capture the likely complexity between years and month on month differences within those years. So we’ll settle for an interaction depth level of 3.

Now for alpha. After some quick research on the internet, we see that droughts are a huge cost on the Australian economy (http://www.economist.com/node/9065059). So, we can imagine that over-predicting average rain is significantly more costly than under-predicting rain. If farmers are expecting more rain and thus adequate precautions aren’t taken (perhaps additional irrigation, purchasing of insurance, etc.), the economic toll could be devastating and ripple effects of high food prices felt throughout the economy. So we will assume policymakers consider overestimates to be more costly than underestimates. Intuitively, the cost of a farmer being over prepared is significantly less than the cost of a farmer being caught unawares and losing all his crops in expectation of more rain. We’ll settle for a cost ratio of 3:1, which would make our alpha equal to 0.25.

Thus we run GBM with the parameters set as above. The best number of iterations is 8228. We can see that there’s overfitting, as the green line (our average out of sample performance) decreases much less quickly than the black line (our in sample performance).

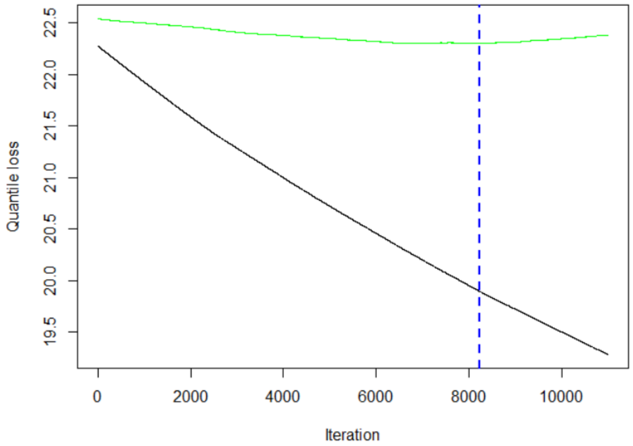


Figure 5

We regress the actual values of the response against our fitted values to get some understanding of how they do. The red line is the 45 degree line.

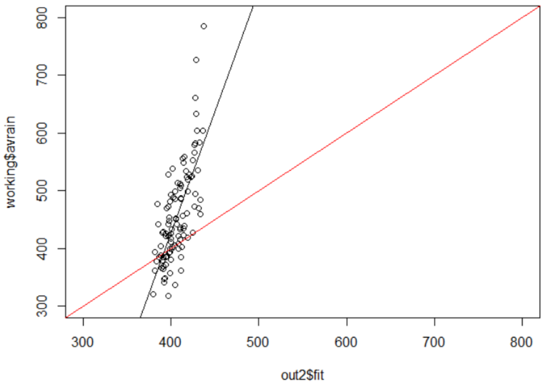
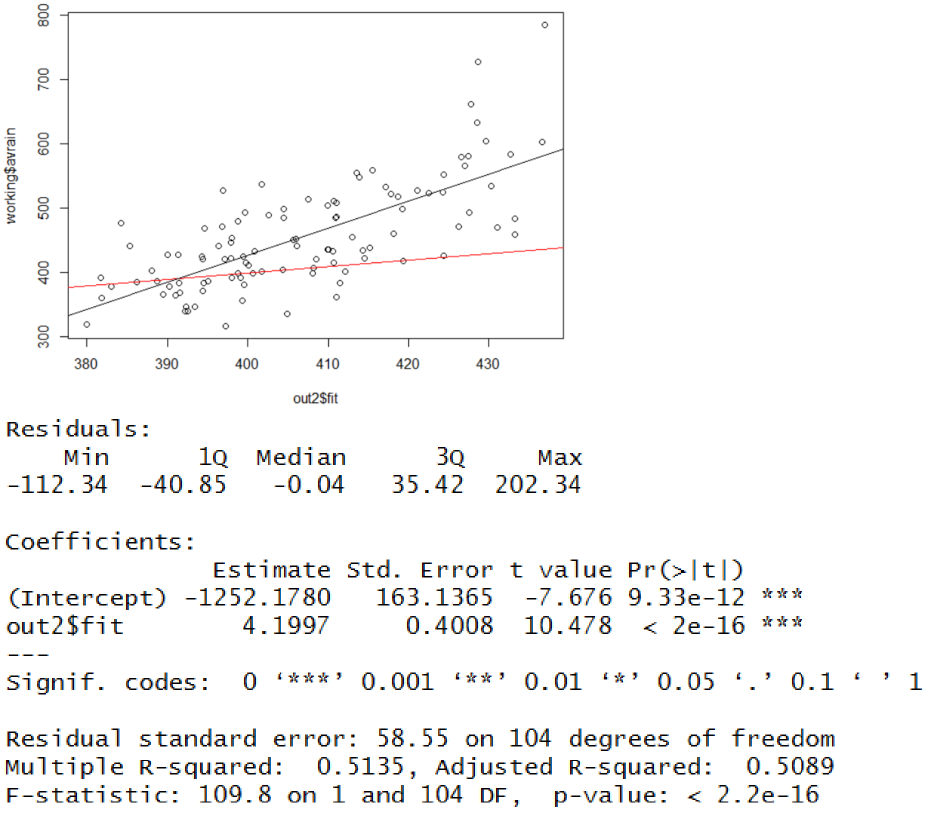


Figure 6

As we can see, most of our data falls above the red line. Considering this is a plot of actual values vs. fitted ones, this is good—we set an alpha of .25 to estimate the 25th percentile, which means we wanted about 3 underestimates for every overestimate. And indeed, the above plot demonstrates that most of our actual values fall above the 45 degree line. This means we underestimated most of them, which is just what we wanted.

Here is a closer look and a summary of the OLS regression of actual values on fitted:



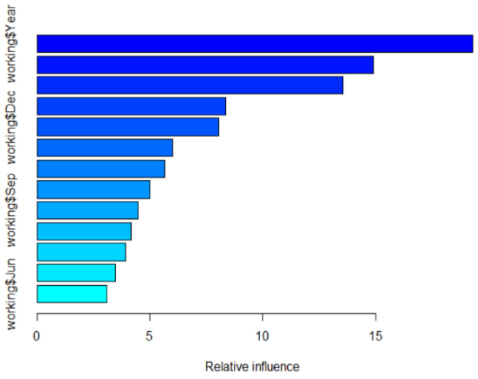
```
Residuals:
    Min      1Q  Median      3Q     Max 
-112.34  -40.85   -0.04   35.42  202.34 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1252.1780   163.1365  -7.676 9.33e-12 ***
out2$fit        4.1997     0.4008  10.478  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 58.55 on 104 degrees of freedom
Multiple R-squared:  0.5135,	Adjusted R-squared:  0.5089 
F-statistic: 109.8 on 1 and 104 DF,  p-value: < 2.2e-16
```

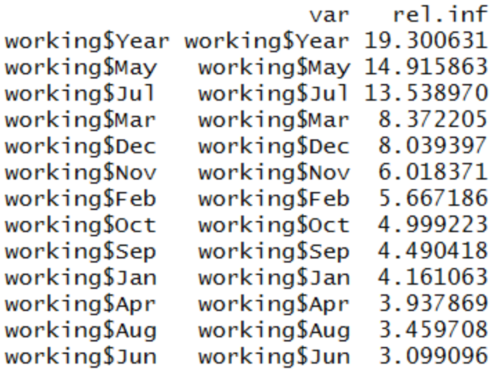
Figure 7

As instructed by the GBM code template, we will look at the distribution of residuals from this regression to determine fit. This makes sense, as we are seeing how well our predicted values match up to our response. Note that we can only use the r squared value as an eyeball estimate, since it’s very difficult to determine the number of degrees of freedom that we had. Thus, taken with a grain of salt, an r-squared of .5 suggests we're doing alright. Our residuals are almost centered at 0, (median of -.04), a good sign. We can think of the OLS line as evaluating fit taking into account our cost ratio. So, when we see a max residual of 202.4 and a minimum residual of -112, we know that our fitting attempt is not doing well in the tails. We are not as worried about the under-predictions, but the over-predictions even with a 3:1 cost ratio are what matter (meaning the residual of -112).

Variable Importance Plots

As Berk explains on page 274 of *Statistical Learning*, the GBM package records the amount that deviance is reduced each time a predictor is chosen for to split the data, and the average decrease in deviance over trees as a proportion of the total decrease from all predictors is reported.





```
                        var   rel.inf
working$Year working$Year 19.300631
working$May   working$May 14.915863
working$Jul   working$Jul 13.538970
working$Mar   working$Mar  8.372205
working$Dec   working$Dec  8.039397
working$Nov   working$Nov  6.018371
working$Feb   working$Feb  5.667186
working$Oct   working$Oct  4.999223
working$Sep   working$Sep  4.490418
working$Jan   working$Jan  4.161063
working$Apr   working$Apr  3.937869
working$Aug   working$Aug  3.459708
working$Jun   working$Jun  3.099096
```

Figure 8

The variable with the highest relative importance is year. This is plausible, as any multi-year trends will certainly have be captured by the year term more effectively than by the averages of any individual month. As the global climate has probably changed in the last 100 years, this makes sense. It is interesting to see May and December in the top 5 predictors, because they are roughly the end and beginning of the El Niño/La Niña period. This suggests our fitting procedure is capturing El Niño effects, which is good, because those with subject matter knowledge would likely say it matters. As for why July and March are also among the 5 most important variables, it's not clear. July is the middle of the Australian winter, and March is the beginning of fall. This might suggest that our procedure is picking up on seasonal effects, which is also good, as they likely matter too.

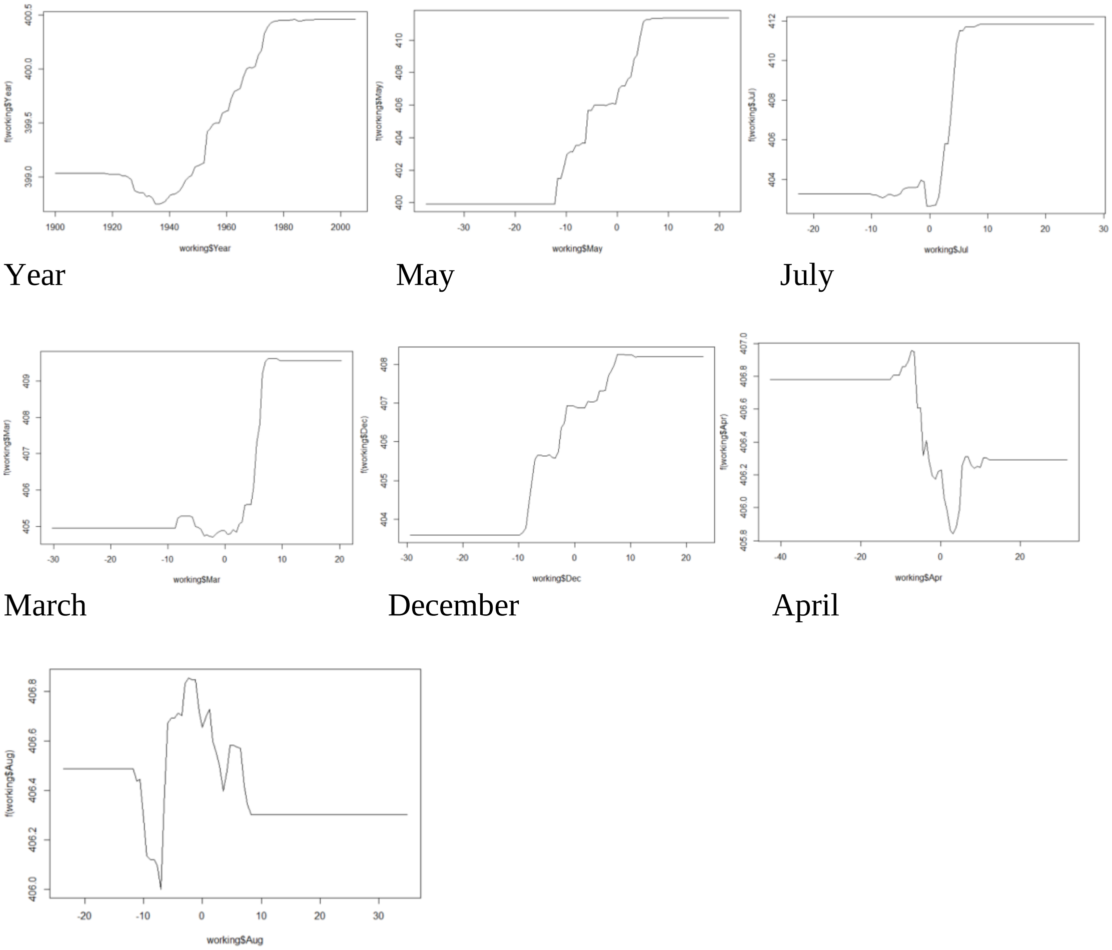
It is also worth noting that, as discussed earlier, the significant collinearity between the predictors will likely confound the inference that we try to do with this plot. Since the algorithm is choosing variables which, when split on, reduce deviance in a node the most, the collinear predictors that are slightly weaker will often not get chosen. Thus they won't be split on as often, and will seem irrelevant. For example, it's not clear why February should be any more significant than January—it's possible that the collinearity between the two is part of this.

While there are complications, we have some reason to believe that our procedure is capturing that the beginning and ending months of El Niño matter, seasonal trends, as well as the yearly trends. Each of these have good subject matter reason to be related to the response.

Partial Dependence Plots

For a given predictor x, these are generated by taking a unique value of x (call it c) and setting all values of predictor x equal to c. Then the algorithm runs, and records the predictions, and takes the mean value. It does this for each unique value of x, and finally plots the predicted values from this procedure against the unique values of predictor x, so we can maybe get some intuition about how variable x is related to the response, holding all else constant.

As per the instructions, displayed below are the “plots the matter”—we include the variables that were the most important according to the variable importance plot, as well as two anomalous months.



Year

May

July

March

December

April

August

Figure 9

We see that the five variables that mattered the most all have an interesting trend: low values of SOI predict the same low amount of rain, then there is some steep increase between predicted average rain and SOI levels, and finally the plot levels off to high SOI values consistently predicting the same high average rain value. The fact that high SOI values correspond to lots of rain and low SOI values to low amounts of rain is good, as subject matter expertise claims the same thing. (http://www.weatherzone.com.au/climate/indicator_enso.jsp?c=soi). It is interesting that April and August buck this trend (they are the only two months to do so). April, again, is the end of the El Niño/La Niña period, so perhaps that is signal our model is capturing (but again, it is not clear why January, the beginning of this period, has a trend like all the others). For August, we are even less sure. While it is the end of the Australian winter, the ends of the other three seasons all exhibit the typical trend.

It is also important to think algorithmically here. We used the default minimum bucket size in stochastic gradient boosting of 10. This setting, combined with a small data set means it's more difficult for our regression trees to grow large and complicated. It's likely that, for each month, our classifier is putting observations past a certain SOI threshold all in one terminal node—hence the exact same fitted values for high and low SOI values within the same month. If there's more granular signal amongst the tail values of SOI and average rain, with our limited data and need to protect against overfitting, our fitting procedure can't pick it up. Perhaps if we had more data, we might get different predicted values in the tails, and the partial plots would look different.

It's interesting that year also exhibits the logistic-like curve. Perhaps in the 1900s the climate was radically different, then between 1920 and 1960 human industrial activity and other factors radically altered some aspect of climate that affects rainfall, and that effect has since plateaued. Again, as non-climate scientists it's difficult for us to really know.

As the confounders above make apparent, we have to take our results with a grain of salt. Though it is interesting that our fitting procedure does find that between the high and low "plateau" values of average rainfall there is a period where the slope of average rainfall against SOI is sharply positive.

Conclusions:

Most of our SOI values were concentrated between a "normal" range of -10 to +10, with some outliers. Our predictors were correlated, particularly El Niño/La Niña months to themselves and non-El Niño/La Niña months to themselves.

We then considered that because of Australia's reliance on agriculture, overestimates probably cost the economy more than underestimates. Having quantile regression trees as our base classifier allowed us to take this asymmetric cost ratio into account. Choosing an alpha of 0.25 (or in effect, estimating the median with overestimates 3 times as costly as underestimates) we implemented stochastic gradient boosting.

It is worthwhile to remember that Professor Berk said in class that there are not many well developed tools to assess boosting's performance when we don't have test data. So we settle for looking at the residuals when the predicted is regressed on the fitted values. It was nice to see the residuals of this line essentially centered around 0. However, our RMSE of 58 was not great. When we compared our results to the 45 degree line, we saw that we seemed to be able to estimate the 25th percentile well, which was encouraging. We also learned some interesting things about how the predictors were related to the response—El Niño/La Niña months seem to matter, and as far as we can tell rain levels are relatively constant for more extreme SOI values but change much more quickly in the middle. The same phenomenon occurring for year might suggest that there were large scale climate trends across the globe that our model found. 106 observations is small for any sort of machine learning problem, and the predictors in this dataset were quite correlated. In this setting, it seems stochastic gradient boosting performed reasonably well.