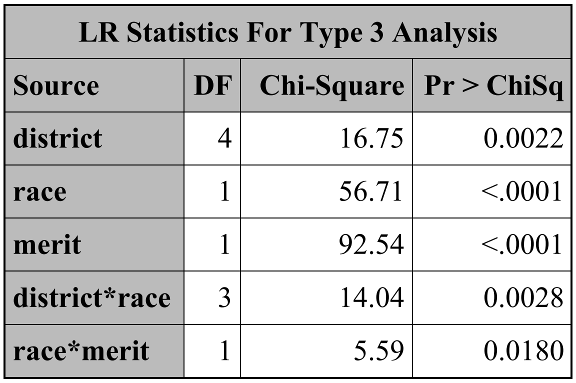
Matt Isaac
Rates and Proportions
Homework 3
Due 10/27/17

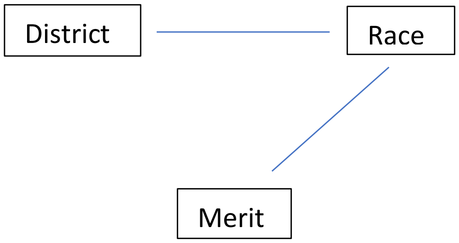
1. Data from problem 4.15
    a. District, Merit Pay, and Race are not mutually independent. I fit a log-linear model with all three variables and no interactions. The deviance was 23.4216 on 13 degrees of freedom. $P(X^2_{13} > 23.4216) = 0.0369$. This is < .05, thus this model does not adequately describe these data. This indicates that District, Merit Pay, and Race are not mutually independent. It will require some interactions to obtain an adequate model.
    b. Fitting the homogeneous associations model gave us a scaled deviance of 3.1525 on 6 degrees of freedom. $P(X^2_6 > 3.1525) = 0.789$. This is > .05, this this model does indeed adequately describe these data.
    c. After removing non-significant terms from the homogeneous associations model, we were left with a final model containing the terms in the table below:



LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
district	4	16.75	0.0022
race	1	56.71	<.0001
merit	1	92.54	<.0001
district*race	3	14.04	0.0028
race*merit	1	5.59	0.0180

The deviance for this model was 3.6853 on 9 degrees of freedom. $P(X^2_9 > 3.6853) = 0.931$. This indicates that this model does adequately describe these data. I chose to stop removing terms at this point because everything left is still statistically significant, and our current model is adequate for these data.

    d. Independence Graph:



No, Race and Merit are not conditionally independent given district. However, District and Merit ARE conditionally independent.

e. After fitting a logit model (Merit$_{yes}$/Total = District Race), I used PROC GENMOD to obtain the deviance to determine whether this model adequately describes these data. The deviance was 1.9464 on 4 degrees of freedom. P(X^2_6 > 1.9464) = 0.9245. Thus, this model does adequately describe these data.

f. The only non-significant terms (out of the two terms in the model) is district (p = 0.9584). After removing district, our model has a deviance of 2.5867 on 8 degrees of freedom. P(X^2_8 > 2.5867) = 0.9575. Thus, this mode adequately describes these data. The model only has one term (race) left in it. See table below:

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
race	1	8.08	0.0045

g. Black workers are only about half (0.447 times) as likely to get Merit Pay as white workers. (Odds ratio obtained from PROC LOGISTIC)

2. GHQ Data

a. A psychotropic drug is “any drug capable of affecting the mind, emotions, or behavior.” (as defined on https://www.medicinenet.com/script/main/art.asp?articlekey=30807)

b. I first fit the model with all two-way interactions. This was adequate to describe the data, so I went on to remove the non-significant terms.
I noticed that all of the two-way interactions were highly non-significant, so I decided to try removing all of them at once. The resulting model was adequate to describe these data, with a deviance of 14.3087 on 13 degrees of freedom. P(X^2_{13} > 14.3087) = 0.3525. All the terms left in the model are statistically significant, so I decided to use this model as my final model. The terms in the final model are displayed in the table below:

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
gender	1	45.46	<.0001
age	4	173.53	<.0001
score	1	234.98	<.0001

c. The terms in our model tell us that all three variables – age group, gender, and GHQ score have a significant effect on the likelihood of a person taking psychotropic drugs. I I will discuss each variable and the effect it has.
Females are 1.873 times as likely to use psychotropic drugs as men are. That is almost twice as likely! I found this to very surprising and interesting.
It seems that people in age groups 16 to 29, 30 to 44, and 45 to 64 are all quite a bit less likely to take psychotropic drugs as people in age group 75 and over. 16 to 29 year-olds are 0.183 times as likely to take psychotropic drugs as people 75 and older. Likewise, people ages 30 to 44 and 45 to 64 are respectively 0.394 and 0.678 times as likely to

take psychotropic drugs as people 75 and older. The trend we see here is that as age increases, so does the likelihood of taking psychotropic drugs. Perhaps more meaningful interpretations of these would come if we take the reciprocal of these odds ratios. People ages 75 and up are 1/0.183 = 5.64 times as likely to take psychotropic drugs as 16 to 29 year-olds; 1/0.394 = 2.54 times as likely to take psychotropic drugs as 30 to 44 year-olds, and 1/.678 = 1.47 times as likely to take psychotropic drugs as 45 to 60 year-olds.
Lastly, people with a high GHQ score are a 4.11 times as likely to use psychotropic drugs as people who have low GHQ scores.
I think that it would be safe to conclude that those who are most at risk for psychotropic drug use would be females over 75 who have high GHQ scores.

3. Cancer knowledge data
   a. I used PROC LOGISTIC and backwards elimination to select variables for my final model. I used a cutoff of 0.05 for the significance level to stay. Below is a table summarizing the variable selection process, listing the order in which terms were removed from the model.

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	news*lect*radi*readi	1	14	0.9921	0.3192
2	newspa*lecture*radio	1	13	0.1368	0.7115
3	newspa*radio*reading	1	12	0.1842	0.6678
4	newspaper*radio	1	11	0.0018	0.9662
5	lectur*radio*reading	1	10	0.4890	0.4844
6	radio*reading	1	9	0.0039	0.9504
7	newspa*lectur*readin	1	8	0.9524	0.3291
8	lecture*radio	1	7	1.4374	0.2306
9	newspaper*reading	1	6	2.7920	0.0947
10	newspaper*lecture	1	5	3.0759	0.0795
11	lecture*reading	1	4	3.3601	0.0668

The following terms were retained in the final model: newspaper, lecture, radio, and reading. These terms and their significance levels are displayed in the table below:

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
newspaper	1	31.6944	<.0001
lecture	1	4.8450	0.0277
radio	1	6.4400	0.0112
reading	1	78.4285	<.0001

b. The odds ratios given by SAS were “no vs yes” for each of the information sources. I have taken the reciprocal of these odds ratios to give us the likelihood of having good information if the subject does use a given information source (“yes vs no”). To begin with, reading seems to be most associated with the largest increase in good knowledge. Those who read are 2.667 times as likely to have good information as those who do not read. The newspaper is the next most effective source of information. Those who read the newspaper are 1.91 times as likely to have good information about cancer as those who do not read the newspaper. These two are unsurprising to me. It seems that if things are published or printed, they go through a more serious review process. Another possible explanation would be that people who are taking time to read are more educated and more serious about retaining knowledge, whereas people listen to the radio for entertainment, not just education. Lectures seem to be next on the list of most increasing good knowledge about cancer. People who attend lectures are 1.522 times as likely to have good knowledge about cancer as those who do not attend. And lastly, people who listen to the radio are 1.36 times as likely to have good information about cancer as those who do not listen to the radio.

   It is interesting (and comforting) to note that all of these sources tend to increase “good information” about cancer. None of them have a negative impact on obtaining good knowledge. Books seem to be the most associated with the largest increase in good knowledge, followed by newspapers, lectures, and the radio.

4. Lava Bed data
   a. The requested model is a poor to moderate fit to these data. For logistic regression, we use r-square as a metric for how well our model fits. In this case, the Max-rescaled R-Square = 0.3818. Another indication of this moderate fit is that the AUC computed from the ROC curve is 0.8295.
   b. *Skip*
   c. I used backwards elimination to select variables for this model. Below is a table summarizing the backwards elimination process. 11 of the original 29 variables were removed from the model. 0.05 was used as a the significance level to stay in the model.

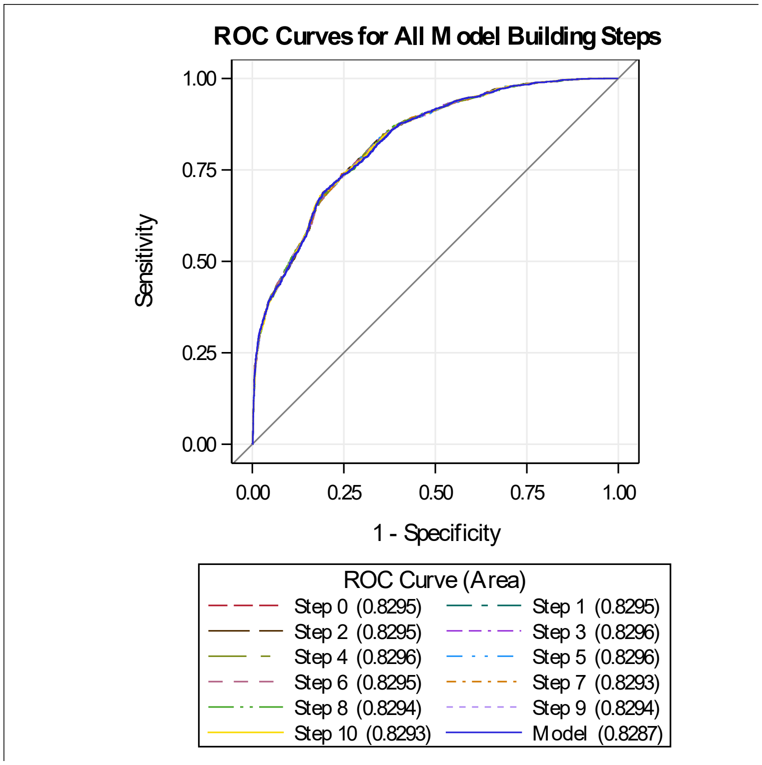
Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	MaxTempAve	1	26	0.0141	0.9056
2	RelHumidAve	1	25	0.0449	0.8322
3	MoistIndexDiff	1	24	0.4081	0.5229
4	VapPressDefAve	1	23	0.4919	0.4831
5	DayTempDiff	1	22	0.8881	0.3460
6	MinTempAve	1	21	1.3435	0.2464
7	PotGlobRadAve	1	20	2.3340	0.1266
8	MaxTempDiff	1	19	2.6630	0.1027
9	MoistIndexAve	1	18	2.6149	0.1059
10	EvapoTransAve	1	17	0.7684	0.3807
11	TransAspect	1	16	2.5969	0.1071

The following table displays the variables that remain in the final model.

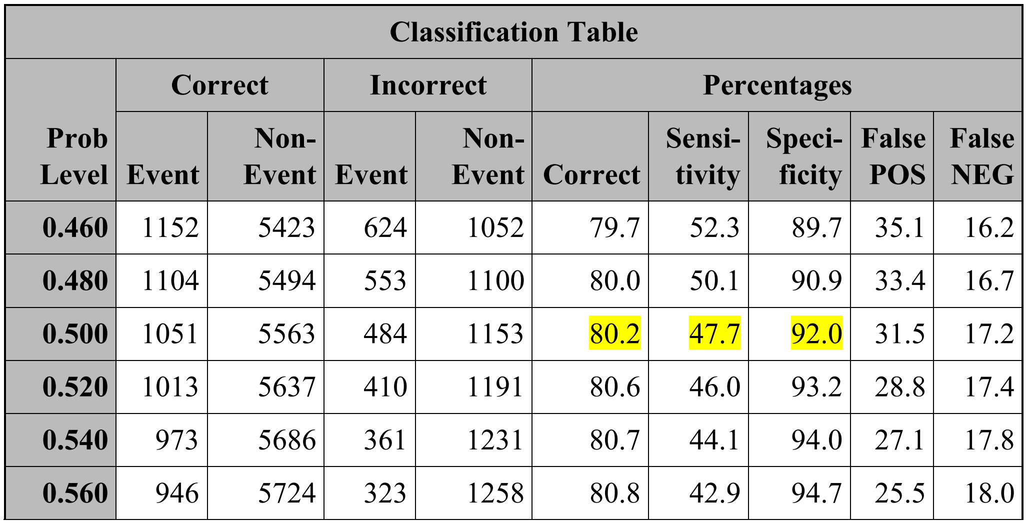
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	62.7399	132.9	0.2229	0.6369
DistRoadTrail	1	-0.00225	0.000143	245.9152	<.0001
PercentSlope	1	-0.1876	0.0206	82.6796	<.0001
DegreeDays	1	0.00758	0.00109	47.9682	<.0001
EvapoTransDiff	1	-71.5640	9.0342	62.7493	<.0001
PrecipAve	1	10.7122	2.1676	24.4219	<.0001
PrecipDiff	1	12.0393	3.2300	13.8931	0.0002
RelHumidDiff	1	28.3241	7.9459	12.7066	0.0004
PotGlobRadDiff	1	-0.0385	0.00961	16.0704	<.0001
AveTempAve	1	-47.4072	7.0236	45.5583	<.0001
AveTempDiff	1	40.6318	7.9116	26.3758	<.0001
DayTempAve	1	34.7253	7.7611	20.0193	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
MinTempDiff	1	-41.0765	7.5222	29.8194	<.0001
VapPressDefDiff	1	39.0335	10.1817	14.6973	0.0001
SatVapPressAve	1	-25.2923	5.9566	18.0295	<.0001
SatVapPressDiff	1	30.4294	11.0108	7.6374	0.0057
Elevation	1	-0.0149	0.00259	32.8349	<.0001

The ROC curves and AUC for each curve were also calculated for each model in backwards elimination process. Each curve and the respective AUC is displayed in the graph below.



d. There are several metrics that we can use to determine the predictive power of this final model. First, the ROC curve gives us an idea of how accurately the models are able to predict. It is notable that the ROC curves for each step in the model selection process are essentially right on top of each other. This tells us that removing those 11 variables did not decrease the predictive power of our model. This is confirmed by the AUC of our starting and final models. We started with an AUC of 0.8295 and after removing 11 variables selected a model with an AUC of 0.8287. This is technically a decrease in predictive power, but the difference between the starting and ending AUC's is insignificant. An AUC of 0.8287 describes a moderate fit of our data. It is fair, but not outstanding. Lastly, we can use the Max-Rescaled R-Square as a metric for model predictive power. The R-Square of 0.3799 leads me to believe that this model isn't quite as good as the AUC made it sound. Ultimately, the highest PCC that is possible is about 80.8 % (see table segment below). This gives us a sensitivity of 41.5 % and specificity of 95.1 %.

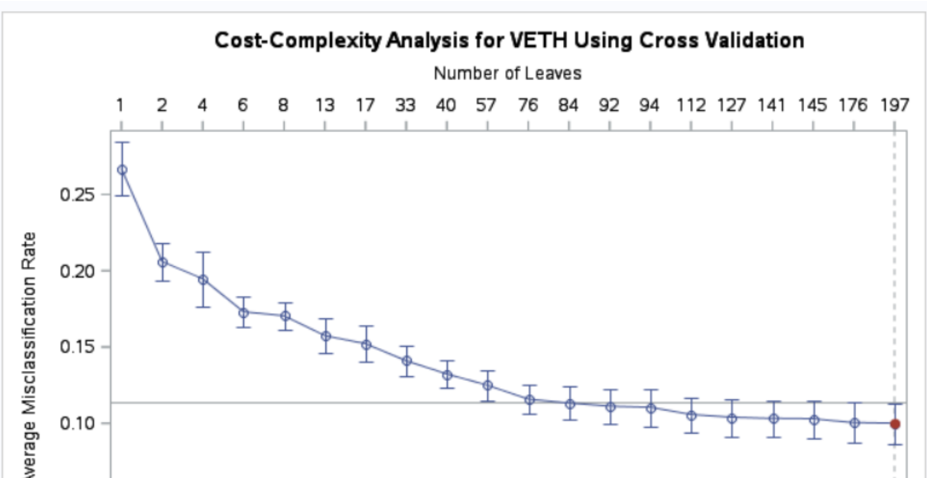


Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensitivity	Specificity	False POS	False NEG
0.460	1152	5423	624	1052	79.7	52.3	89.7	35.1	16.2
0.480	1104	5494	553	1100	80.0	50.1	90.9	33.4	16.7
0.500	1051	5563	484	1153	80.2	47.7	92.0	31.5	17.2
0.520	1013	5637	410	1191	80.6	46.0	93.2	28.8	17.4
0.540	973	5686	361	1231	80.7	44.1	94.0	27.1	17.8
0.560	946	5724	323	1258	80.8	42.9	94.7	25.5	18.0

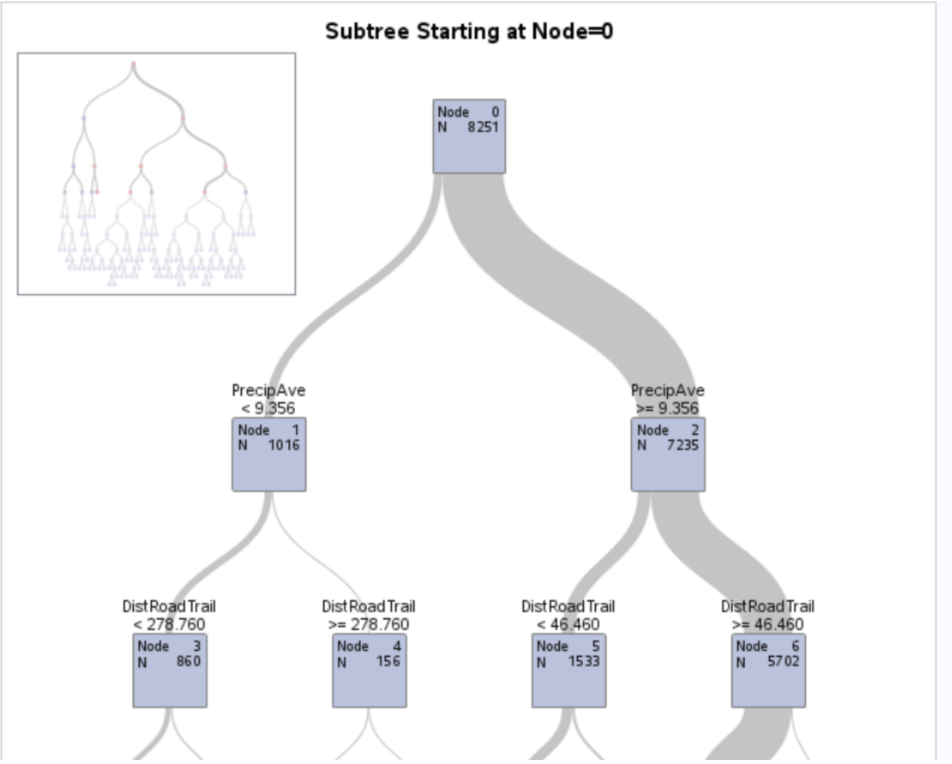
e. DistRoad is estimated to have a parameter of -0.00028 and is highly statistically significant (p < 0.0001). The negative parameter tells us that sites closer to roads are more closely associated with absences of mullein.
DistTrail, on the other hand, is estimated to have a parameter of 0.000833, and is highly statistically significant (p < 0.0001). This indicates that sites closer to trails are more closely associated with presences of mullein.
DistRoadTrail is estimated to have a model parameter of -0.00225, and is highly statistically significant (p < 0.0001). Because this is negative, it indicates that the sites closer to trails or roads are more closely associated with absences than the sites further from roads or trails.
I would interpret this to mean that mullein seeds aren't transported into the are from cars as much as they are on boots and shoes of hikers.

f. I first fit the full (unpruned) classification tree to these data. The cp plot is shown below:



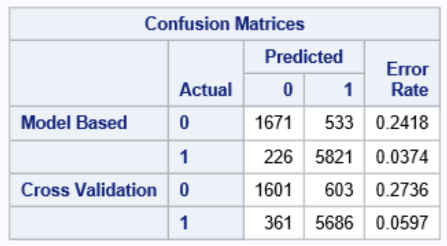
From this plot (keeping in mind the 1-SE rule of thumb) I decided to prune tree back to have 76 nodes. The average misclassification rate for the 76-node tree does fall slightly outside of one standard error of the minimum misclassification rate, but it is insignificantly different from the misclassification rate of the 84-node tree. Choosing the 76-node tree gives us a simpler tree essentially the same misclassification rate. I then refit the tree, pruning back to 76 leaf nodes.

g. Displayed below are the first few nodes of the classification tree:
Beginning with node 0, the first variable that we split on is PrecipAve. Those sites that received average precipitation < 9.356 cm (or inches?) were sent to the left (Node 1), and the sites that received >= 9.356 cm (or inches) of precipitation were sent to the right (Node 2). Node 1 and Node 2 were then split on DistRoadTrail. From Node 1 (the sites with < 9.365 cm of precip), the sites < 278.760 units of measure from a road or trail were sent to the left (Node 3) and the sites >= 278.760 units of measure from a road or trail were sent to the right (Node 4). From Node 2 (the sites with >= 0.365 cm of precip), the sites < 46.460 units of measure from a road or trail were sent to the left (Node 5) and the sites >= 46.460 units of measure from a road or trail were sent to the right (Node 6). So it appears that some of the most important variables for predicting the presence/absence of mullein are average precipitation and distance from a road/trail.



h. Recall that the percent correctly classified from the final logistic regression model was 80.8 %, with a sensitivity of 41.5 % and specificity of 95.1 %. The cross-validated accuracy of the classification tree is (1601 + 5686 / 8251) * 100% = 88.32 %. The sensitivity is 94.03 % and specificity is 72.64 %. In all respects, the classification tree has a higher predictive power than the logistic regression model. The PCC is higher, and the sensitivity and specificity are both higher. Regardless of the researchers specific needs, I think that they can be better met by modeling presences and absences with the classification tree.

Displayed below is the confusion matrix for the classification tree.



Confusion Matrices

		Predicted		Error
	Actual	0	1	Rate
Model Based	0	1671	533	0.2418
	1	226	5821	0.0374
Cross Validation	0	1601	603	0.2736
	1	361	5686	0.0597