Exam 1: Categorical Data Analysis

Elizabeth Grimes
June 30, 2020

1 Introduction

In this paper, we will answer the following questions in preparation for completing our term project.

1. Write your response variable as a categorical variable and do Chi-Squared Tests of Independence or Testing Independence for Ordinal Data, and calculate the odds ratio.

2. Apply logistic regression to your data, backward variable selection, and do Model Checking using the deviance (LRT and AIC). Report your best logistic regression model.

3. Apply probit regression to your data, backward variable selection, and do Model Checking using the deviance (LRT and AIC). Report your best probit regression model.

4. Compare your best logistic regression model and the best probit regression model with AUC and AIC to find the final model. Report and interpret your final regression model.

2 Background

There are 130 college football teams in the NCAA Division I College Football Subdivision (abbreviated FBS), and their records from the 2019 college football season are the data that we examine in this project. Nearly all of these teams are members of one of 10 FBS conferences. The regular college football season consists of 12 games (except for the University of Hawaii, which plays 13 games, and all those schools that play against the University of Hawaii). Teams that win at least 6 regular season games qualify to participate in one of 40 bowl games, taking place during December and January. In addition, each conference holds a conference championship game after the regular season ends. Two bowl games each year are designated as College Football Playoff games. Teams are ranked by the playoff selection committee throughout the season, and after all regular season and conference championship games are played, the top 4 teams are chosen to participate in the two College Football Playoff games (which are two of the 40 bowl games). The winners of those two games advance to the College Football Playoff Championship game. As a result, teams may play anywhere from 12 to 15 games during a season (barring game cancellation for weather or other extenuating circumstances). During the 2019 season, all FBS teams played at least 12 games.

The ten FBS conferences are divided into two groups, which is relevant to our study. The Power 5 conferences (abbreviated P5) are the Atlantic Coast Conference (ACC), the Southeastern Conference (SEC), the Big Ten, the Big 12, and the Pac-12. These conferences are considered to include the best college football teams in the country, and since the College Football Playoff was implemented in 2014, every team that qualified for the playoff was a member of one of the Power 5 conferences. The remaining conferences are known unofficially as the Group of 5. The schools in these conferences are generally assumed to play less challenging schedules and to be not

as competitive. The independent schools are not members of any conference. Their strength of schedule varies widely, and while some are perceived as highly competitive and elite teams, others are not.

3 Question 1: Response Variables and Chi-Square Test

The response variable is the number of wins achieved by each college football team in 2019. This variable can be written in a number of different ways: as an ordinal variable ranging from 0 to 15, the number of teams who achieved winning seasons (more than 50% win percentage), the number of teams who became bowl-eligible during the season (winning at least 6 games), or the number of teams who won 10 games or more (a commonly accepted metric for a highly successful football season).

We perform a chi-square test for independence on the number of wins (winning record, bowl qualification, and 10-win seasons) and the conference. Frequency tables are provided for each chi-square test. None of these tests showed that the variables were independent, and because of the small number of teams in each conference, these tests do not necessarily produce an accurate result. For the variables winning record and conference, the chi-square test for independence produces a test statistic of $\chi^2 = 1.0847$, with 10 degrees of freedom and a p-value of 0.9998. Thus, winning record and conference are independent. For the variables of bowl qualification and conference, the chi-square test for independence produces a test statistic of $\chi^2 = 2.4815$, with 10 degrees of freedom and a p-value of 0.9911. Thus, bowl qualification and conference are independent. For the variables 10-win season and conference, the chi-square test for independence produces a test statistic of $\chi^2 = 6.8861$, with 10 degrees of freedom and a p-value of 0.7361. Thus, achieving a 10-win season and conference are independent.

Because most of these categories have fewer than 5 data points, it is worth collapsing the conferences into larger groups. We consider each of the win metrics for the teams in the Power 5 and Group of 5. Because there are so few FBS Independent schools, we include these with the Group of 5 schools. Five of these six schools would not be considered P5 schools. Notre Dame is the exception–it has an affiliation with the ACC, a Power 5 conference, and would rightly be considered a P5 school, but for ease of analysis we group all FBS Independent schools with the Group of 5. These data are displayed in Tables 4, 5, and 6.

A chi-square test for independence was performed between the variables of winning record and P5 vs. Group of 5 teams. This test produced a test statistic of $\chi^2 = 0.00018319$, with 1 degree of freedom and a p-value of 0.9892. A chi-square test for independence was performed between bowl qualification and P5 vs. Group of 5 teams, producing a test statistic of $\chi^2 = 0.58187$, with 1 degree of freedom and a p-value of 0.4456. A chi-square test for independence was performed between a 10-win season and P5 vs. Group of 5 teams, producing a test statistic of $\chi^2 = 0.30657$, with 1 degree of freedom and a p-value of 0.5798. Thus, all three measures of number of wins were independent from membership in a P5 or Group of 5 conference.

4 Question 2: Logistic Regression Model Generation

Our data contains more than 150 different measures of offensive, defensive, and special teams play. It is prohibitively complicated to include all of these measures in our logistic regression. The list of measures can be roughly divided into categories of offensive measures, defensive measures, special teams measures, miscellaneous measures (those that do not reflect specifically on offense or defense), and response measures. There are 49 measures for offense, 58 measures for defense,

Conference	Yes	No
AAC	7	5
ACC	7	7
Big 12	7	4
Big Ten	7	6
C-USA	7	7
FBS Independent	3	3
MAC	6	6
Mountain West	7	5
Pac-12	6	6
SEC	8	6
Sun Belt	5	5
Total	70	60

Table 1: Winning records for FBS football teams in 2019, by conference

Conference	Yes	No
AAC	7	5
ACC	10	4
Big 12	7	4
Big Ten	8	5
C-USA	8	6
FBS Independent	3	3
MAC	8	4
Mountain West	7	5
Pac-12	7	5
SEC	10	4
Sun Belt	5	5
Total	80	50

Table 2: Bowl qualifications for FBS football teams in 2019, by conference

Conference	Yes	No
AAC	3	9
ACC	1	13
Big 12	2	9
Big Ten	3	10
C-USA	1	13
FBS Independent	1	5
MAC	0	12
Mountain West	2	10
Pac-12	2	10
SEC	4	10
Sun Belt	2	8
Total	21	109

Table 3: Number of teams achieving 10 wins in 2019, by conference

Group	Yes	No
Power 5	35	35
Group of 5	29	31
Total	64	66

Table 4: Winning records for FBS football teams in 2019, by P5 vs. Group of 5

Group	Yes	No
Power 5	42	38
Group of 5	22	28
Total	66	64

Table 5: Bowl qualification for FBS football teams in 2019, by P5 vs. Group of 5

Group	Yes	No
Power 5	12	52
Group of 5	9	57
Total	21	109

Table 6: Ten win seasons for FBS football teams in 2019, by P5 vs. Group of 5

26 measures for special teams, 10 miscellaneous measures, and 7 response measures. Of these, 22 different measures rank the 130 teams from top to bottom on their performance. For ease of calculation for the purposes of this midterm, we will use only these rankings. As we complete this investigation, we will examine the other pieces of data and determine the variables that will be most beneficial for our model.

First, we define a subset of data that incorporates all of the ranking data (22 measures). We then perform logistic regression on these 22 measures. We use winning records, bowl qualification, and ten-win seasons as the response variables. Of these, the ten-win season does not produce a convergent algorithm, so we discard this option. For the purposes of this midterm, we will consider only whether a team has a winning record or not as the response variable. Future analyses may consider whether the model will be different for bowl qualification.

4.1 Winning Record

Table 7 displays the coefficients for the logistic regression using the 22 rankings as explanatory variables and whether the team has a winning record as the response variable. This model has an AIC of 89.524.

Table 7: Coefficients for Logistic Regression, Winning Record vs. RankData, Version 1

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	30.976241	12.341147	2.51	0.0121
Off.Rank	0.0189	0.043937	0.43	0.6671
Def.Rank	-0.466169	0.210716	-2.212	0.0269
Penalty.Rank	-0.015487	0.01644	-0.942	0.3462
First.Down.Rank	-0.11932	0.060913	-1.959	0.0501
First.Down.Def.Rank	0.160427	0.071388	2.247	0.0246
X4th.Down.Rank	-0.029774	0.027468	-1.084	0.2784
X4rd.Down.Def.Rank	-0.047664	0.028738	-1.659	0.0972
Kickoff.Return.Rank	-0.032056	0.022445	-1.428	0.1532
Kickoff.Return.Def.Rank	-0.043459	0.025112	-1.731	0.0835
Passing.Off.Rank	-0.037307	0.046753	-0.798	0.4249
Pass.Def.Rank	0.116135	0.077865	1.491	0.1358
Punt.Return.Rank	-0.05062	0.023901	-2.118	0.0342
Punt.Return.Def.Rank	-0.012255	0.018365	-0.667	0.5046
Redzone.Off.Rank	-0.022936	0.023745	-0.966	0.3341
Redzone.Def.Rank	-0.037333	0.020328	-1.837	0.0663
Rushing.Off.Rank	-0.03699	0.036949	-1.001	0.3168
Rushing.Def.Rank	0.138086	0.093995	1.469	0.1418
Sack.Rank	-0.104504	0.061365	-1.703	0.0886
Scoring.Off.Rank	0.083527	0.056562	1.477	0.1397
Scoring.Def.Rank	0.141458	0.07718	1.833	0.0668
Tackle.for.Loss.Rank	0.04211	0.03739	1.126	0.2601
X3rd.Down.Rank	-0.002781	0.024816	-0.112	0.9108
Time.of.Possession.Rank	-0.05722	0.031861	-1.796	0.0725
Turnover.Rank	-0.054828	0.037006	-1.482	0.1384

Next, we use the "drop1" command and the likelihood ratio test to determine which variables should be removed from the model. We find that Rushing.Off.Rank has an AIC of 87.593, an LRT

statistic of 0.0685, and a p-value of 0.79354, which is the highest of all variables in the model. Thus we remove Rushing.Off.Rank. The coefficients for the updated model are displayed in Table 8.

Table 8: Coefficients for Logistic Regression, Winning Record vs. RankData, Version 2

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	21.563305	8.407864	2.565	0.0103
Def.Rank	-0.351833	0.171538	-2.051	0.0403
First.Down.Def.Rank	0.170442	0.065995	2.583	0.0098
Kickoff.Return.Def.Rank	-0.037935	0.019225	-1.973	0.0485
Kickoff.Return.Rank	-0.02305	0.01734	-1.329	0.1838
Off.Rank	-0.044307	0.033471	-1.324	0.1856
Pass.Def.Rank	0.057205	0.062578	0.914	0.3606
Penalty.Rank	-0.023707	0.015955	-1.486	0.1373
Punt.Return.Def.Rank	-0.005244	0.012811	-0.409	0.6823
Punt.Return.Rank	-0.030651	0.016411	-1.868	0.0618
Redzone.Def.Rank	-0.021883	0.015184	-1.441	0.1495
Redzone.Off.Rank	-0.010377	0.017101	-0.607	0.544
Rushing.Def.Rank	0.074538	0.07465	0.999	0.318
Sack.Rank	-0.071298	0.042336	-1.684	0.0922
Scoring.Def.Rank	0.093531	0.06458	1.448	0.1475
Scoring.Off.Rank	0.011822	0.034952	0.338	0.7352
Tackle.for.Loss.Rank	0.033503	0.028805	1.163	0.2448
Time.of.Possession.Rank	-0.052734	0.026078	-2.022	0.0432
Turnover.Rank	-0.033458	0.025485	-1.313	0.1892
X4rd.Down.Def.Rank	-0.028089	0.01788	-1.571	0.1162
X3rd.Down.Rank	-0.008459	0.017994	-0.47	0.6383
X4th.Down.Rank	-0.021257	0.01932	-1.1	0.2712

Performing the "drop1" command again, we find that the variable Scoring.Off.Rank has an AIC of 85.707, a LRT statistic of 0.1145, and a p-value of 0.73507, the highest remaining p-value. We remove this variable next.

Our 4th version of the logistic regression model drops the variable Punt.Return.Def.Rank, with an AIC of 83.826, a LRT statistic of 0.1183, and a p-value of 0.730842.

Our 5th version of the logistic regression model drops the variable X3rd.Down.Rank, with an AIC of 81.975, a LRT statistic of 0.1488, and a p-value of 0.699661.

Our 6th version of the logistic regression model drops the variable Redzone.Off.Rank, with an AIC of 80.482, a LRT statistic of 0.5078, and a p-value of 0.476087.

Our 7th version of the logistic regression model drops the variable Pass.Def.Rank, with an AIC of 79.048, a LRT statistic of 0.5655, and a p-value of 0.4520447.

Our 8th version of the logistic regression model drops the variable Rushing.Def.Rank, with an AIC of 77.287, a LRT statistic of 0.2394, and a p-value of 0.6246459.

Our 9th version of the logistic regression model drops the variable Tackle.for.Loss.Rank, with an AIC of 76.689, a LRT statistic of 1.4018, and a p-value of 0.2364245.

Our 10th version of the logistic regression model drops the variable Turnover.Rank, with an AIC of 76.245, a LRT statistic of 1.5555, and a p-value of 0.2123213.

Our 11th version of the logistic regression model drops the variable Scoring.Def.Rank, with an AIC of 75.131, a LRT statistic of 0.8865, and a p-value of 0.3464341.

Our 12th version of the logistic regression model drops the variable Kickoff.Return.Rank, with an AIC of 74.010, a LRT statistic of 0.8793, and a p-value of 0.3483873.

Our 13th version of the logistic regression model drops the variable X4th.Down.Rank, with an AIC of 73.966, a LRT statistic of 1.956, and a p-value of 0.1619635.

Our 14th version of the logistic regression model drops the variable X4rd.Down.Def.Rank, with an AIC of 73.594, a LRT statistic of 1.6281, and a p-value of 0.2019653.

Our 15th version of the logistic regression model drops the variable Redzone.Def.Rank, with an AIC of 73.111, a LRT statistic of 1.517, and a p-value of 0.2180756.

Our 16th version of the logistic regression drops the variable Penalty.Rank, with an AIC of 72.279, a LRT statistic of 1.168, and a p-value of 0.2799150.

Our 17th version of the logistic regression drops the variable Punt.Return.Rank, with an AIC of 71.658, a LRT statistic of 1.379, and a p-value of 0.2402650.

Subsequent versions of the logistic regression model cause the AIC to go back up, so the 17th version is the best fit. The coefficients for this model and the included variables are listed in Table 9. The residual deviance of the model is 57.658 on 123 degrees of freedom, and the AIC is 71.658. All variables except for Sack.Rank are significant at at least the 0.05 level.

Table 9: Coefficients for Logistic Regression, Winning Record vs. RankData, Version 17

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	9.12132	1.89904	4.803	1.56E-06
Def.Rank	-0.11685	0.02778	-4.206	2.60E-05
First.Down.Def.Rank	0.10396	0.0266	3.908	9.31E-05
Kickoff.Return.Def.Rank	-0.02381	0.01004	-2.37	0.01778
Off.Rank	-0.03492	0.01107	-3.154	0.00161
Sack.Rank	-0.01762	0.01151	-1.531	0.12572
Time.of.Possession.Rank	-0.0436	0.01444	-3.02	0.00253

Of the 6 remaining explanatory variables, 4 (Def.Rank, First.Down.Def.Rank, Kickoff.Return.Def.Rank, and Sack.Rank) are considered to be ranks of a team's ability on defense. Thus it may be true that, as stated by legendary football coach Bear Bryant, "Defense wins championships."

5 Question 3: Probit Regression Model Generation

We utilize the same subset of data that we did for the logistic regression–the 22 rankings of various aspects of performance throughout the season. In order to compare with the logistic regression, we use winning record as the response variable. The coefficients for the initial probit regression model are shown in Table 10.

For the purpose of brevity, the variables that are dropped from each version of the probit model are displayed in Table 11. The AIC varies from 87.712 to 70.559.

As in the logistic regression model, there are 6 explanatory variables in the probit regression model. The coefficients for the model are displayed in Table 12. The model has a residual deviance of 56.559 on 123 degrees of freedom, and an AIC of 70.559. All explanatory variables are the same as those in the logistic regression.

Table 10: Coefficients for Probit Regression, Winning Record vs RankData, Version 1

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	11.15562	4.322177	2.581	0.00985
Def.Rank	-0.199357	0.090975	-2.191	0.02843
First.Down.Def.Rank	0.092894	0.034447	2.697	0.007
Kickoff.Return.Def.Rank	-0.02035	0.010161	-2.003	0.0452
Kickoff.Return.Rank	-0.011826	0.009287	-1.273	0.20289
Off.Rank	-0.022361	0.018383	-1.216	0.22384
Pass.Def.Rank	0.036291	0.034276	1.059	0.2897
Penalty.Rank	-0.011314	0.008757	-1.292	0.19636
Punt.Return.Def.Rank	-0.002074	0.006926	-0.299	0.76463
Punt.Return.Rank	-0.016484	0.008825	-1.868	0.06179
Redzone.Def.Rank	-0.012753	0.008309	-1.535	0.12481
Redzone.Off.Rank	-0.005202	0.009869	-0.527	0.59812
Rushing.Def.Rank	0.042482	0.04008	1.06	0.28917
Rushing.Off.Rank	-0.002806	0.008657	-0.324	0.74586
Sack.Rank	-0.038373	0.021818	-1.759	0.07862
Scoring.Def.Rank	0.052223	0.034242	1.525	0.12723
Scoring.Off.Rank	0.006395	0.018705	0.342	0.73244
Tackle.for.Loss.Rank	0.019613	0.015375	1.276	0.20209
Time.of.Possession.Rank	-0.027556	0.013567	-2.031	0.04225
Turnover.Rank	-0.015789	0.014933	-1.057	0.29035
X4rd.Down.Def.Rank	-0.013856	0.009822	-1.411	0.15834
X3rd.Down.Rank	-0.006583	0.011041	-0.596	0.55105
X4th.Down.Rank	-0.009711	0.010273	-0.945	0.34449

Table 11: Variables Dropped from Probit Regression Model

Version	Variable to Delete	Df	Deviance	AIC	LRT	P-value
2	Rushing.Off.Rank	1	43.712	87.712	0.0971	0.755392
3	Punt.Return.Def.Rank	1	43.797	85.797	0.0856	0.769813
4	Scoring.Off.Rank	1	43.901	83.901	0.1041	0.747004
5	X3rd.Down.Rank	1	44.089	82.089	0.1878	0.664759
6	Redzone.Off.Rank	1	44.697	80.697	0.6079	0.435593
7	Pass.Def.Rank	1	45.368	79.368	0.6713	0.412595
8	Rushing.Def.Rank	1	45.431	77.431	0.0632	0.8015366
9	Tackle.for.Loss.Rank	1	47.078	77.078	1.6465	0.1994381
10	Turnover.Rank	1	48.479	76.479	1.4016	0.2364578
11	Scoring.Def.Rank	1	49.188	75.188	0.7086	0.3999162
12	Kickoff.Return.Rank	1	49.931	73.931	0.7433	0.3886144
13	X4th.Down.Rank	1	51.577	73.577	1.646	0.1995408
14	X4rd.Down.Def.Rank	1	52.83	72.83	1.2526	0.2630607
15	Redzone.Def.Rank	1	54.311	72.311	1.4816	0.2235259
16	Penalty.Rank	1	55.526	71.526	1.214	0.2704788
17	Punt.Return.Rank	1	56.559	70.559	1.033	0.309448

Table 12: Coefficients for Probit Regression, Winning Record vs. RankData, Version 17

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	5.274227	1.008353	5.231	1.69E-07
Def.Rank	-0.067222	0.014792	-4.544	5.51E-06
First.Down.Def.Rank	0.05896	0.014015	4.207	2.59E-05
Kickoff.Return.Def.Rank	-0.013419	0.005559	-2.414	0.01578
Off.Rank	-0.019747	0.006072	-3.252	0.00115
Sack.Rank	-0.010479	0.006463	-1.622	0.10489
Time.of.Possession.Rank	-0.024756	0.007755	-3.192	0.00141

6 Question 4: Model Comparison

To compare the models, we examine the AIC for each. The final logistic model has an AIC of 71.658 and a residual deviance of 57.658 on 123 degrees of freedom. The final probit model has an AIC of 70.559 and a residual deviance of 56.559 on 123 degrees of freedom. Since our goal is a lower AIC, the probit model is better for this data. However, we will need to generate ROC curves and calculate AUC to determine whether each curve is truly a good fit for the data.

Figure 1 shows the ROC curve for the logistic regression. The AUC is 0.970, indicating that the logistic regression is an excellent predictor of whether a football team will have a winning record based on their ranks on the six measures.

Figure 1: ROC curve for logistic regression

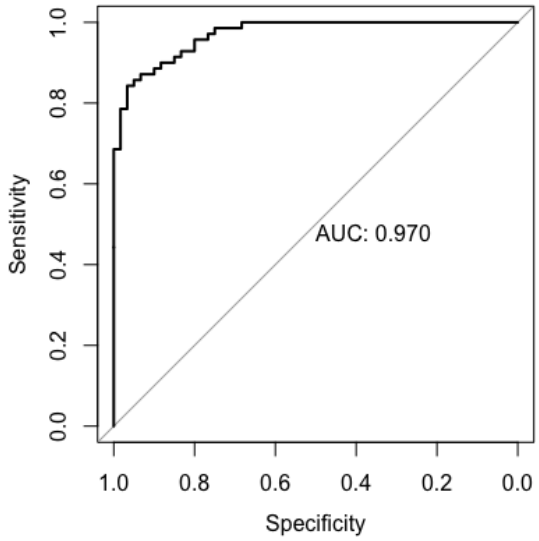
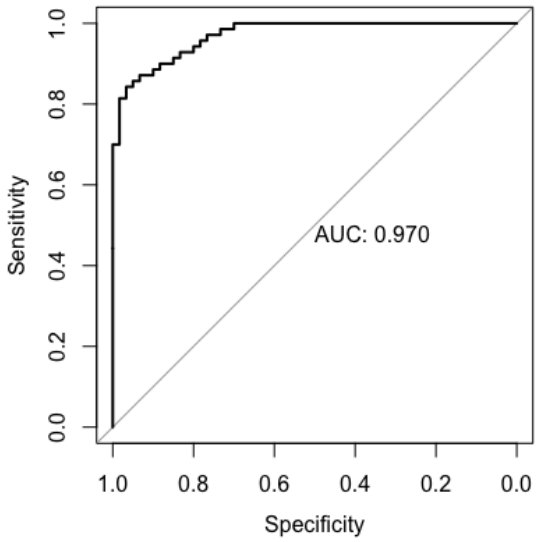


Figure 2 shows the ROC curve for the probit regression. The AUC is 0.970 again, though the curve is slightly different for the probit regression. The probit regression is also an excellent predictor of whether a football team will have a winning record based on their ranks on the six measures.

Figure 2: ROC curve for probit regression



7 Conclusion

We have generated two excellent models to predict whether a college football team will have a winning record, based on their ranks on 6 different measures: overall offense, overall defense, first down defense, kickoff return defense, number of sacks, and time of possession. We have found that the probit regression model is slightly better, because its AIC is lower, though both models have the same AUC. Future expansions of this investigation could investigate which of the many other measures can predict winning record, as well as investigating whether different variables should be used to generate models to predict whether teams qualify for a bowl game or whether teams attain a ten win season. We also have not investigated interactions between the various variables, and those should also be investigated. Finally, this data covers only the 130 FBS football teams for the 2019 season. Having expanded data for additional seasons would enable us to improve the model even further. The R code for data preparation and each question is attached to this document.