What Can the Cowboys Do to Win?

Will Hallgren

November, 2021

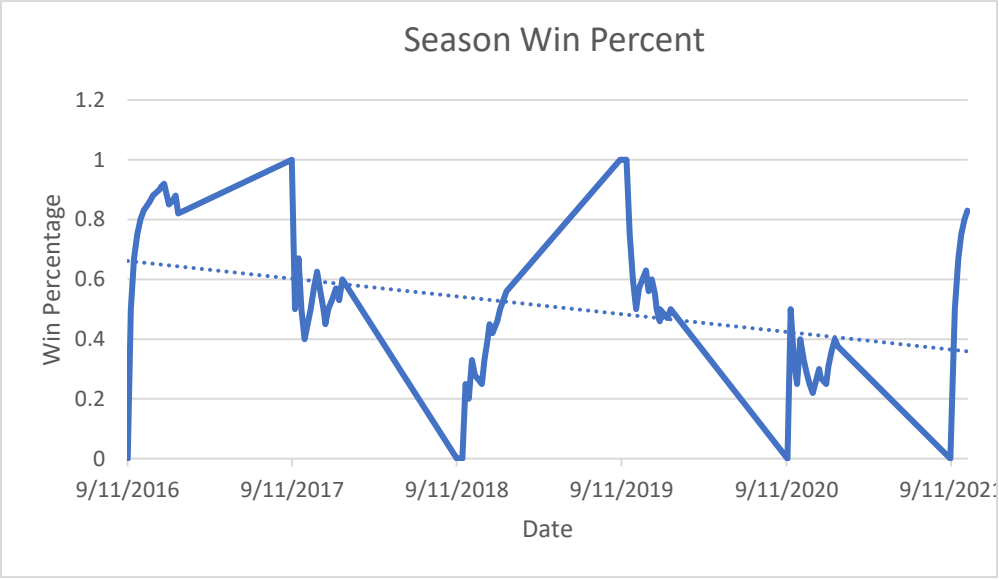
ECON 5670: Applied Econometrics

University of North Texas

Introduction

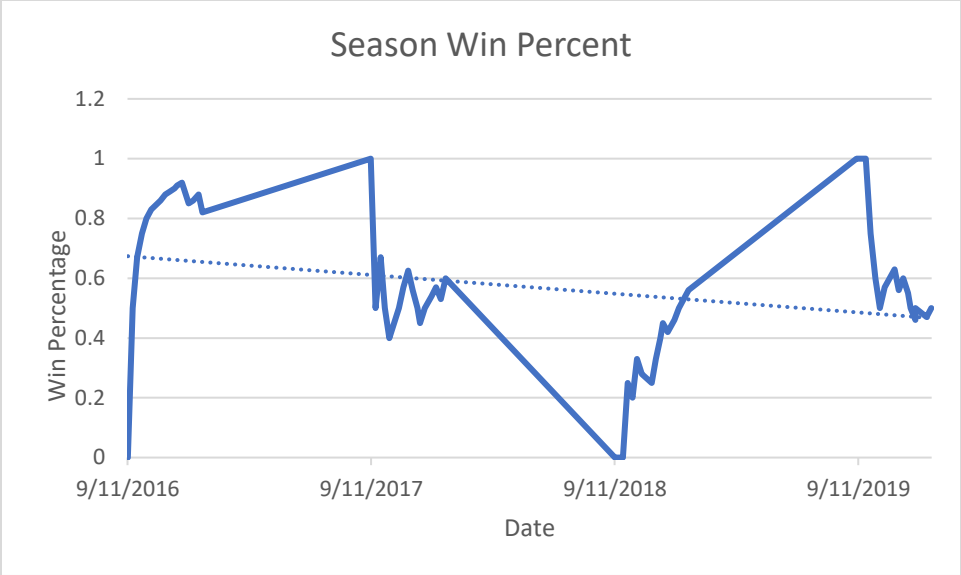
The Dallas Cowboys, a National Football League team owned by Jerry Jones, are the most valuable sports franchise in the world, even when compared to sports teams from different sports, from different countries. The Cowboys consistently sell over 10,000 more tickets than there are seats that are available at AT&T stadium for every home game. Economic theory would tell us that without competitive balancing measures taken by the NFL, just based on the size of the franchise, the Cowboys should win nearly every game. They would have the most funds to spend on coaches, star players, athletic trainers, and facilities. With such a loyal fanbase, the Cowboys have the largest budget of any sports team in the world. Thanks to regulation by the NFL, there is more variation in the outcomes of games. I believe, however, that the Cowboys, being the largest sports franchise in the world, would still have a competitive edge even with considerable competitive balancing. The Cowboys have been on a downward trend, especially considering the 2020 season.

Graph 1: Season Win Percent from 2016-2020



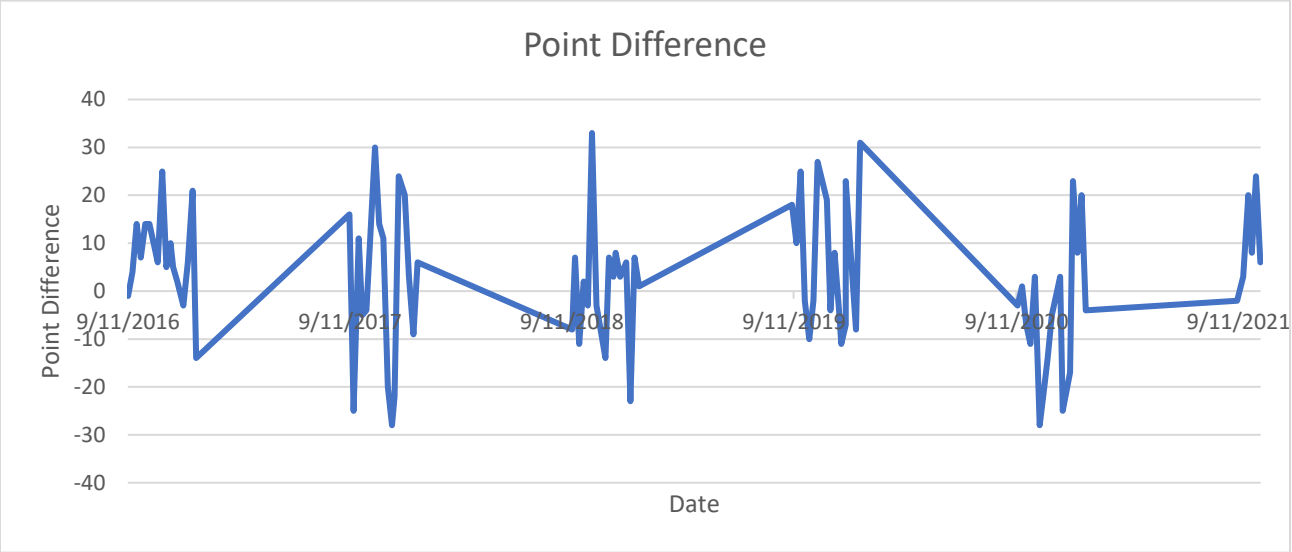
The 2021 season seems to be more promising. The Cowboys currently have an 85.7%-win percentage for the season. However, excluding the 2020 season, because of the pandemic and because their starting quarterback was injured for the season, there was still a downward trend from 2016 to 2019. These were seasons that I expected to be relatively successful.

Graph 2: Win Percent for Season from 2016-2019 Season



There is something amiss about this trend. The Cowboys' management must be overlooking some critical component to the game if the team with the most resources is on a downward trend since recruiting a quarterback with a relatively, consistently high quarterback rating. There are other positions and star players that I believe influence the outcome of the game however the quarterback seems to have the greatest effect. The Cowboys also lost by more in the 2017, 2018, and 2020 seasons than in 2016. Graph 3 shows the difference between points scored and points allowed, indicating how much the Cowboys won or lost by each game. The spread widens in 2017, 2018, and 2020 from 2016. Games in those years were more likely to be landslide victories or losses. Both 2016 and 2019 appear to be relatively good seasons for the Cowboys, and the spread appears to be smaller.

Graph 3: Difference Between Points Scored and Points Allowed



In this study, I wish to find the factors that contribute to the outcome of a Cowboys game primarily because I am curious to see if there is something that Jerry Jones is overlooking or if the NFL is truly sufficient at balancing the league. The team is on a downward trend, and I believe that something must be done to improve the current trajectory of the team's winning ability. I would like to know if there is anything that can be done by the team management to improve the odds of winning a game, or if the outcome is random enough that you might as well flip a coin to predict a win or a loss. I would also like to understand why the spread between points scored and points allowed is so large in seasons with more losses. In addition, if this model can indeed accurately predict the outcome of a game, then the model could potentially be a useful tool for sports betting. Potentially, other teams could also benefit from such an analysis and football could become more competitive overall.

The Data

This data comes from all the regular season games starting in 2016 until the most recent game in 2021. I found this data on a website called "Football Database," (https://www.footballdb.com/teams/nfl/dallas-cowboys/history). I chose to start in 2016 because that was the year that the Cowboys drafted their current star quarterback, Dak Prescott. In total, there are 86 games observed. The dependent variable in this data set is "$game\_won_i$". It is a dummy variable that equals 1 if the game was won and it equals 0 if the game was lost.

There are 20 independent variables that I collected for this data set. "$win\_percent_i$" is the percentage of games won for the season for game "i". I start the calculations over at the beginning of every season. I expect this variable to have a positive coefficient because a higher win percentage indicates that the team has been winning in the past. This could be because of some external factor such as a winning team composition or successful training, or it could be mental as well. A team that has been winning is likely to be more confident in their ability to win, which may improve a team's odds of winning. "p_scored_i" indicates the points scored by the Cowboys in game "i". I expect this variable to have a positive coefficient because the more a team scores, the more likely they are to win. "$p\_allowed_i$" indicates the points scored by the opposing team in game "i". I expect this variable to have a negative coefficient because the more the opposing team scores, the more likely the Cowboys are to lose. "$home\_dum_i$" is a dummy variable that equals 1 if the game was played at the Cowboys' home stadium, AT&T stadium in game "i". I expect this variable to have a positive coefficient because teams typically have an advantage playing games at their home stadium. "$q\_rate\_Cowboys_i$" is the quarterback rating for the quarterback with the greatest number of attempts for the Cowboys for game "i".

Typically, this variable is Dak Prescott's quarterback rating but in 2020, Prescott was injured early in the season and was unable to play again for the rest of the season. I expect this variable to have a positive coefficient because the performance of the quarterback relates directly to a team's ability to score. The quarterback also acts as an on-field strategist. The better the quarterback plays, the more points a team is likely to score. "$q\_rate\_opp_i$" is the quarterback rating for the quarterback with the greatest number of attempts for the opposing team in game "i". I expect this variable to have a negative coefficient because the better the opposing team's quarterback plays, the more likely they are to score. "$yards\_pass_i$" is the net number of yards gained by passing the ball for the Cowboys in game "i". I expect this variable to have a positive coefficient because gaining yards contributes to scoring points. "$yards\_pass\_opp_i$" is the net number of yards gained by passing the ball for the opposing team in game "i". I expect the variable to have a negative coefficient because the opposing team gaining yards leads to them scoring, which would make wins less likely. "$yards\_rush_i$" is the number of yards gained by the Cowboys from rushing the ball in game "i". I expect this variable to have a positive coefficient for the same reason that the passing variable is positive. More yards gained leads to more points scored which would increase the chance of winning. "yards_rush_opp" is the number of yards gained by the opposing team from rushing the ball in game "i". I expect this variable to have a negative coefficient because the more yards gained by the opposing team, the more points they score, which would decrease the chance of winning. "$num\_penalty_i$" indicates the number of penalties received by the Cowboys in game "i". I expect this variable to have a negative coefficient because penalties result in yard penalties which would hinder the Cowboys' ability to score. "$loss\_penalty_i$" indicates the number yards lost to penalties in game "i". I expect this variable to have a negative coefficient because with more yards lost to penalties, the Cowboys

are less likely to score. "num_fumble$_i$" is the number of fumbles that the Cowboys had in game "i". I expect this variable to have a negative coefficient because fumbling the ball costs the offense a down, even if there are no yards lost, which makes it harder to score in the series. "loss_fumble$_i$" is the number of yards lost to fumbles in game "i". I expect this variable to have a negative variable coefficient because more yards lost to fumbles would have a negative impact on the Cowboys' ability to score. "num_sack$_i$" is the number of times the quarterback was sacked in game "i". I expect this variable to have a negative coefficient because getting sacked costs the team a down and usually includes a heavy yard loss. "loss_sack$_i$" is the number of yards lost to getting sacked in game "i". I expect this variable to have a negative coefficient because losing yards hinders the team's ability to score. "field_goal$_i$" is the number of field goals scored by the Cowboys in game "i", excluding the extra point attempt after a touchdown. I believe that this variable will have a positive coefficient because field goals contribute to points scored, which would make the Cowboys more likely to win. "halftime_dum$_i$" is a dummy variable that equals 1 if the Cowboys were winning by the halftime and equals 0 if they were tied or losing in game "i". I expect this variable to have a positive coefficient because winning at the half means that the team goes into the next half with a point advantage. Players are likely to play more confidently if they were winning at the half and are equally likely to play more diffidently if they were losing at the half. This affects some players more than others; however, I believe that it does influence overall team performance. "attendance_capacity$_i$" is the ratio between the attendance for the game and the stadium capacity for game "i". I am looking to see if the fullness of the stadium influences the team's performance. I expect this variable to have a positive coefficient because I believe that a fuller stadium has a mental effect on the players that makes them want to win more, even at away games. Stadium attendance also effects a team's

budget. Finally, "percent_total_possession$_i$" is the percentage of time in the whole game that the Cowboys had possession over the ball. I expect this variable to be positive because maintaining control over the ball allows the team to score more and while maintaining control over the ball, the opposing team cannot score.

Table 1: Variable Definitions and Expected Coefficient Signs

Variable Name:	Definition:	Expected Coefficient Sign:
game_won$_i$	= 1 if the game was won in game "i", = 0 if the game was lost	N/A
Win_percent$_i$	Percentage of games won for the season for game "i"	+
P_scored$_i$	Points scored by the Cowboys in game "i"	+
P_allowed$_i$	The points scored by the opposing team in game "i"	-
Home_dum$_i$	= 1 if the game was a home game in game "i", = 0 otherwise	+
q_rate_Cowboys$_i$	The quarterback rating for the quarterback with the greatest number of attempts for the Cowboys for game "i"	+
q_rate_opp$_i$	The quarterback rating for the quarterback with the greatest number of attempts for the opposing team in game "i"	-
Yards_pass$_i$	Net number of yards gained by passing the ball for the Cowboys in game "i"	+
Yards_pass_opp$_i$	The net number of yards gained by passing the ball for the opposing team in game "i"	-
Yards_rush$_i$	The number of yards gained by the Cowboys from rushing the ball in game "i"	+
Yards_rush_opp$_i$	The number of yards gained by the opposing team from rushing the ball in game "i"	-
Num_penalty$_i$	The number of penalties in game "i"	-

Table 1 Continued: Variable Definitions and Expected Coefficient Signs Sample

Variable Name:	Definition:	Expected Coefficient Sign:
Loss_penalty$_i$	The number yards lost to penalties in game "i"	-
Num_fumble$_i$	The number of fumbles in game "i"	-
Loss_fumble$_i$	The number of yards lost to fumbles in game "i"	-
Num_sack$_i$	The number of times the quarterback was sacked in game "i"	-
Loss_sack$_i$	The number of yards lost to getting sacked in game "i"	-
Field_goal$_i$	The number of field goals scored by the Cowboys in game "i"	+
Halftime_dum$_i$	= 1 if the Cowboys were winning by the halftime and = 0 if they were tied or losing in game "i"	+
Attendance_capacity$_i$	The ratio between the attendance for the game and the stadium capacity for game "i"	+
Percent_total_possession$_i$	The percentage of time in the whole game that the Cowboys had possession over the ball	+

Sample Size: There are 86 observations in this study

Table 2: Summary Statistics for All Observations

Variable:	Number of Observations:	Mean:	Standard Deviation:	Minimum:	Maximum:
game_won	86	0.58	0.5	0	1
win_percent	86	0.52	0.26	0	1
p_scored	86	24.98	11.2	0	47
p_allowed	86	22.12	9.37	0	49
home_dum	86	0.5	0.5	0	1
q_rate_Cowboys	86	96.93	27.39	27.5	158.3
q_rate_opp	86	95	19.33	55.3	149.1
yards_pass	86	244.2	90.22	59	481
yards_pass_opp	86	236.49	73.6	60	449
yards_rush	86	136.05	58.03	40	375
yards_rush_opp	86	110.26	57.98	22	323
num_penalty	86	6.42	2.27	2	12
loss_penalty	86	57.09	23.15	10	124
num_fumble	86	1.21	1.08	0	4
loss_fumble	86	0.6	0.74	0	3
num_sack	86	2.33	1.72	0	8
loss_sack	86	15.47	12.64	0	55
field_goal	86	1.77	1.32	0	4
halftime_dum	86	0.53	0.5	0	1
Attendance_Capacity	86	0.9	0.36	0	1.17
percent_total_possession	86	0.51	0.07	0.4	0.77

Table 3: Summary Statistics for games won and games lost

game_won	Number of Obs	Variable	Number of Observations	Mean	Std Dev	Minimum	Maximum
=0	36	win_percent_season	36	0.37	0.23	0	0.85
		p_scored	36	16.22	8.67	0	38
		p_allowed	36	26.56	9.19	10	49
		home_dum	36	0.42	0.5	0	1
		q_rate_Cowboys	36	75.84	21.11	27.5	112.9
		q_rate_opp	36	101.21	17.9	61.2	149.1
		yards_pass	36	233.5	100.88	59	481
		yards_pass_opp	36	218.17	78.4	60	434
		yards_rush	36	100.58	35.68	40	189
		yards_rush_opp	36	137.5	62.49	35	307
		num_penalty	36	6.53	1.92	2	11
		loss_penalty	36	57.44	18.98	10	124
		num_fumb	36	1.22	1.02	0	3
		loss_fumble	36	0.72	0.78	0	2
		num_sack	36	2.92	1.89	0	8
		loss_sack	36	20.75	14.38	0	55
		field_goal	36	1.67	1.35	0	4
		halftime_dum	36	0.25	0.44	0	1
		Attendance_Capacity	36	0.8	0.43	0	1.17
		percent_total_possession	36	0.49	0.07	0.4	0.62
=1	50	win_percent_season	50	0.63	0.22	0.25	1
		p_scored	50	31.28	8.2	6	47
		p_allowed	50	18.92	8.18	0	39
		home_dum	50	0.56	0.5	0	1
		q_rate_Cowboys	50	112.12	20.51	59.5	158.3
		q_rate_opp	50	90.52	19.24	55.3	140
		yards_pass	50	251.9	81.9	93	445
		yards_pass_opp	50	249.68	67.69	111	449
		yards_rush	50	161.58	57.82	51	375
		yards_rush_opp	50	90.64	45.87	22	323
		num_penalty	50	6.34	2.51	2	12
		loss_penalty	50	56.84	25.94	16	115
		num_fumb	50	1.2	1.12	0	4
		loss_fumble	50	0.52	0.71	0	3
		num_sack	50	1.9	1.47	0	7
		loss_sack	50	11.66	9.69	0	40
		field_goal	50	1.84	1.3	0	4
		halftime_dum	50	0.74	0.44	0	1
		Attendance_Capacity	50	0.97	0.29	0	1.17
		percent_total_possession	50	0.53	0.07	0.4	0.77

Table 4: Analysis of home_dum

home_dum	N Obs	Minimum	Maximum
0	**43**	0	1
1	**43**	0	1

Table 5: Analysis of halftime_dum

halftime_dum	N Obs	Minimum	Maximum
0	**40**	0	1
1	**46**	0	1

Table 2 shows the summary statistics for all of the variables for the entire combined data set. Table 3 shows the summary statistics for each of the variables where "game_won$_i$" is zero, and where "game_won$_i$" is one. There are some interesting features of the data that I would like to make note of. The win percentage of the season was higher when more games were won than for games that were not. This reflects that seasons carry momentum. The Cowboys hit winning streaks and losing streaks each season. They seem more likely to lose if they have lost previously and they are more likely to win if they have won previously. In addition, the Cowboys win only slightly more than they lose. There is little spread in average number and number of yards lost to both penalties and fumbles between games lost and games won. There is more spread in both number and loss of yards due to sacks between games lost and games won. The game attendance-capacity ratio was also slightly higher for games that were won than for games that were lost.

Before I can begin regression analysis, I must examine the summary statistics to see if there are any problems with the data. Each variable has a total number of 86 observations, so I

know that there are no missing observations for any of the variables. I see no unreasonable values in the means for each variable. None of the variables have a standard deviation of zero. The minimums and maximums all seem reasonable. I will note that the minimum values for stadium attendance would normally never be zero, however stadium attendance was restricted in 2020 due to the Covid outbreak. For now, I will hold that this observation is reasonable, but I will address this problem further in the future.

There are a few more potential issues with the data that I must address. If the mean for the dependent variable is much smaller than the mean for any regressor, then it will not be possible to compute the estimated coefficient of that regressor because that number will be so small that the computer cannot tell the difference between the true value and zero. The means of the dependent variable and regressors are reasonably close, so I will not need to adjust the measurements for any of the variables. With the dependent variable being a dummy, we must consider the proportion of zeros and ones. If there are too many or too few values where the dependent variable takes a value of one, then there are too many or too few occasions where the trait is not present, or the event does not occur. There need to be enough observations to construe any sort of conclusions from this regression. Ideally there should be 50 percent of observations where "$game\_won_i$" takes on a value of one, but it can be as low as 20 percent and as high as 80 percent before we have reason for concern. There are 50 observations where "$game\_won_i$" takes on a value of 1, out of 86 observations. This is roughly 58 percent of the data, which is very close to ideal. The means for independent dummy variables must be between 0.2 and 0.8 as well because there need to be enough observations in both cases, when the trait is present and when the trait is missing, for the regression results to be accurate. The means for

both independent dummy variables are 0.5 for "home_dum$_i$", and 0.53 for "halftime_dum$_i$". There are no issues with dummy variables in this data set.

When the dependent variable is also a dummy, independent dummy variables must pass the "Dummy Rules" to be considered for a regression. Dummy Rule Zero requires that when the independent dummy variable takes on a value of zero, the dependent variable must take on values of both zero and one. Dummy Rule One requires that when the independent dummy variable takes on a value of one, the dependent variable must take on values of both zero and one. There are two dummy variables that must pass both rules. If the rules are not satisfied, then the independent dummy variable can perfectly predict the dependent variable. Table 4 shows us the number of observations for which "home_dum$_i$" equals one and zero and the minimum and maximum values for the corresponding "game_won$_i$" values. Since the minimums and maximums are zero and one for both outcomes of "home_dum$_i$", the dummy rules are satisfied. Table 5 shows us the number of observations for which "halftime_dum$_i$" equals one and zero and the minimum and maximum values for the corresponding "game_won$_i$" values. Since the minimums and maximums are zero and one for both outcomes of "halftime_dum$_i$", the dummy rules are satisfied for all independent dummy variables.

Finally, the data must also pass the "Largest-Smallest Rules" for continuous independent variables. The Largest-Smallest Inside rule states that the largest value for any independent variable when the dependent variable equals zero must be bigger than the smallest value of the independent variable when the dependent variable equals one. The Largest-Smallest Outside rule states that the largest value for any independent variable when the dependent variable equals one must be bigger than the smallest value of the independent variable when the dependent variable equals zero. If ether rule does not hold, then there exists a threshold value for the

independent variable that can perfectly predict the value of the dependent variable, similar to the dummy rule. The Largest-Smallest Rules are satisfied for all continuous independent variables. I am reasonably confident that there are no issues with this data set.

The Probit Model

The goal of this research is to determine the factors that influence the outcome of a Cowboys game. In this case, I am interested in finding the probability that the Cowboys will win game "i". Many researchers would use the ordinary least squares (OLS) method for data regression however there are some issues with OLS that I would like to avoid. OLS is a linear method of data regression, which may present some issues if the relationship between the dependent and independent variables is non-linear. OLS also has limited reliability in the case of a binary response dependent variable, such as "game_won$_i$". The estimates of the mean and variance in "game_won$_i$" would be biased under OLS. It can make the parameter estimates of the intercept and slopes biased, asymptotically biased, and inconsistent. In addition, the probability estimates for "game_won$_i$" may be outside the range of zero and one. There may be negative probabilities estimated, or probabilities over one hundred percent, which does not provide intuitive interpretation. For this application, I would like to use a method of regression called the probit model. The probit model was first designed by Bliss, C. I. (1934) to find the conditional probability that the dependent variable equals 1.

The first step in the probit model is to define a new indicator dependent variable "game_won$_i$*" on which the data will be regressed.

Table 6: Indicator Dependent Variable Definition

Variable Name:	**Definition:**
Game_won$_i$*	A continuous but unobservable index of the Cowboys' ability to win game$_i$.

The word "probit" comes from combining the words "probability" and "unit". The probit model aims to predict the conditional probability of the dependent variable taking on a value of one, conditioned on some values for the independent variables. From here, I will use general notation until I discuss empirical results, where I call the dependent variable "y_i" and independent variables "x_i". The conditional probability, given by equation 1, is equal to the cumulative density function (CDF) of a standard normal distribution evaluated at $\widehat{Y_\iota}$, the sample regression.

$$(1.) \qquad \widehat{Prob}(y_i = 1|\, x) = \left[1 - F\left(-\widehat{Y_\iota}\right)\right] = F(\widehat{Y_\iota})$$

Where F(*) is the CDF of the normal distribution. The probability of "game_won$_i$" equaling 1 is found by substituting in values for each observation into the estimated regression equation, which gives us $\widehat{Y_\iota}$,then plugging in the resulting $\widehat{Y_\iota}$ into equation 1.

The parameter estimates of the probit model have no direct interpretation because the dependent variable is an unobservable index. Instead, marginal effects for each independent variable are calculated for every observation. The marginal effects are given by equation 2:

$$(2.) \qquad Estimated\ marginal\ effect\ of\ X_{ji} = f\left(\widehat{Y_\iota}\right) \times \left[\frac{\partial(\widehat{Y_\iota})}{\partial X_{ji}}\right]$$

Where f(*) is the probability density function (PDF) of the standard normal distribution.

Absolute goodness of fit is evaluated by a Pseudo-General F Test which is given by the following hypothesis:

(3.) H_0: [$all\ parameters\ except\ the\ traditional\ intercept\ are\ jointly\ equal\ to\ zero$]

vs.

H_a: [$at\ least\ one\ of\ the\ parameters\ except\ the\ tradtional\ intercept\ are\ jointly\ significant$]

The null hypothesis is evaluated using a likelihood ratio (LR) test statistic, which is chi squared distributed. The LR test statistic is given by the following formula:

(4.) $$LR\ test\ statistic = 2 \times \left[ln\, l\left(\hat{\theta}\right) - ln\, l(\tilde{\theta})\right]$$

Where $\ln l\left(\hat{\theta}\right)$ is the maximized value of the log of the likelihood function for the unrestricted regression, or the regression using all independent variables. $\ln l(\tilde{\theta})$ is the maximized value of the log likelihood function for the restricted regression, or the regression that uses only an intercept. The likelihood function comes from the probit model itself. It is the likelihood that "$game\_won_i$" is equal to 1. The natural logarithm is taken in order to simplify the computations. To pass the goodness of fit test, the null hypothesis must be rejected. The LR test statistic must be greater than the critical value for a 90% confidence level with n degrees of freedom, where n is the number of regressors in the model. The number of regressors in the model is the number of restrictions in the null hypothesis. We can also do T-tests to test for significance of individual parameters.

The null hypothesis for a T-test is given by equation 5:

$$(5.) \qquad H_0: \beta_j = 0\ , j = 1, 2, \dots\ , K$$

Vs

$$H_a: \beta_j \neq 0, j = 1, 2, \dots, K$$

Where β_j is the parameter coefficient in the regression equation. The T-test statistic is given by equation 6:

$$(6.) \qquad t = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

Where $\bar{X}$ is the sample mean, μ is the population mean, $\hat{\sigma}$ is the estimated population standard deviation, and n is the number of observations in the sample.

The Hosmer-Lemeshow Test for goodness of fit is another test that evaluates whether the fit of the model is good. It uses the same test statistics as the pseudo-general F tests however the null tests for whether the fit is good. Traditionally the null tests whether the fit is bad. The null hypothesis for Hosmer-Lemeshow is given by equation 7 on the next page:

$$(7.) \qquad H_0: the\ fit\ is\ "good"$$

Vs

$$H_a: the\ fit\ is\ not\ good$$

Finally, we can also take the percentage of correct predictions to evaluate the effectiveness of the model. The probit model gives probabilities that "$game\_won_i$" equals 1. If

the probability given by the model is greater than or equal to 0.5, then the predicted outcome is 1. If the probability given by the model less than 0.5, then the predicted outcome is 0. A probability is calculated for every observation and the predicted outcome is compared to the actual outcome recorded in the data. If the percentage of correctly predicted outcomes for both outcomes is 80% or higher, then the model is considered to be a very good fitting model. We can also test the model's predictive ability by holding out a few observations from the regression and observing if the model can correctly predict their outcomes as well.

Empirical Results