Logistic Regression: What is it and What can I learn from it?

Melodie Rush
Senior Systems Engineer

CUSTOMER LOYALTY TEAM ▪ Support You Can Count On

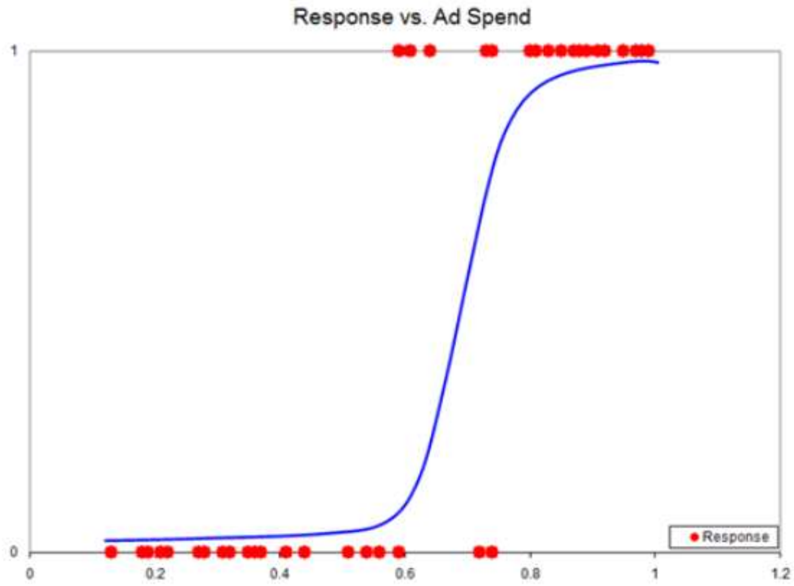
Agenda

- Why would you use it?
  - Goal
  - Application
- What is Logistic Regression?
- Examples
  - Data layout
  - Simple
  - Multiple

What is our goal?

Common Applications

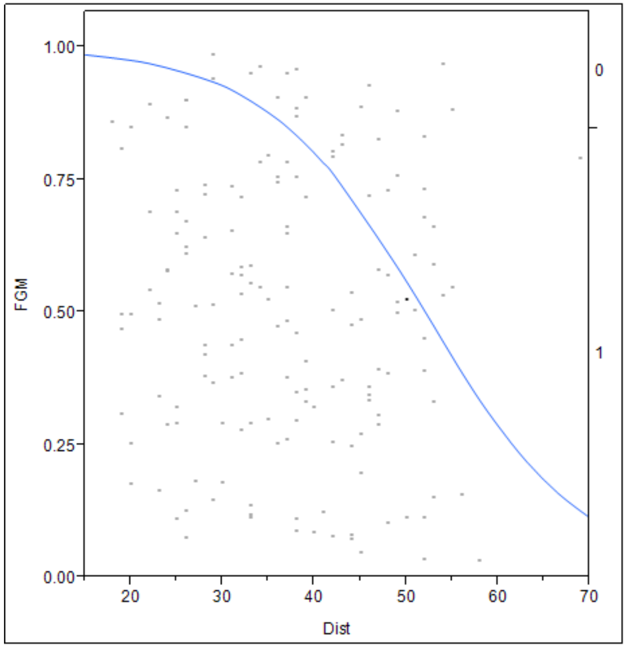
- Target Marketing
- Attrition Prediction
- Credit Scoring
- Fraud Detection
- Customer Satisfaction

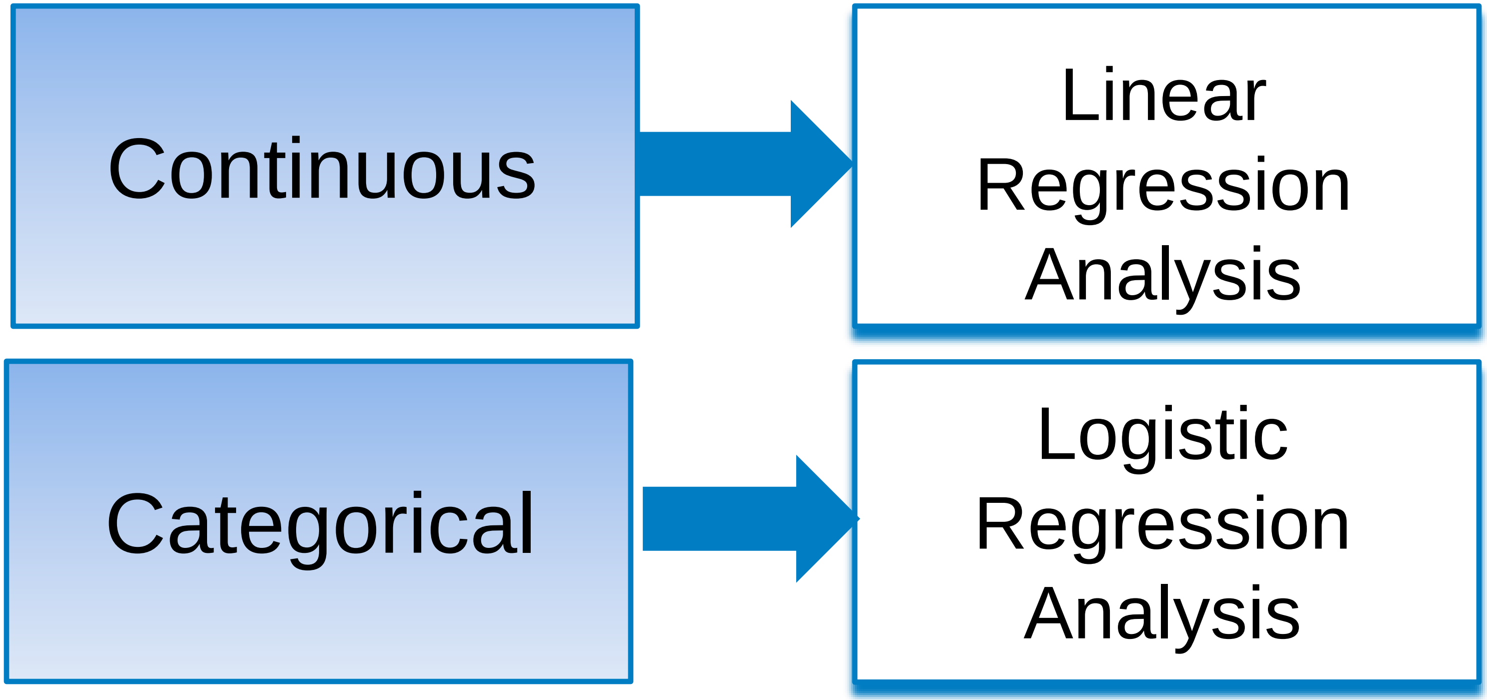


Good or No Good?

What is Logistic Regression?

Logistic Regression is essentially a regression model tailored to fit a categorical dependent variable.





Response

Analysis

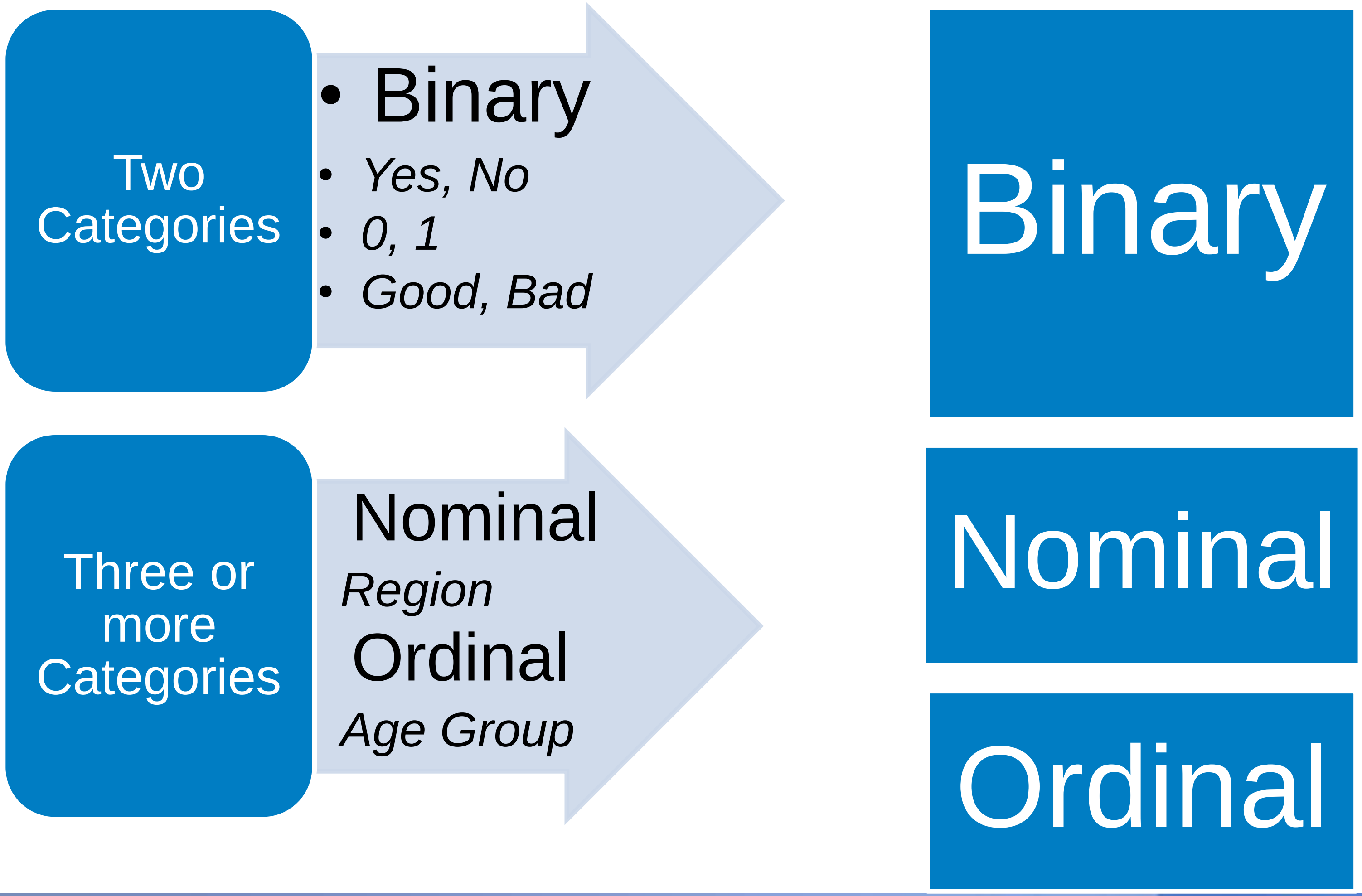
Continuous

Linear Regression Analysis

Categorical

Logistic Regression Analysis

Types of Logistic Regression



Response Variable

Type of Logistic Regression

Two Categories

- Binary
  - *Yes, No*
  - *0, 1*
  - *Good, Bad*

Binary

Three or more Categories

Nominal

Region

Ordinal

Age Group

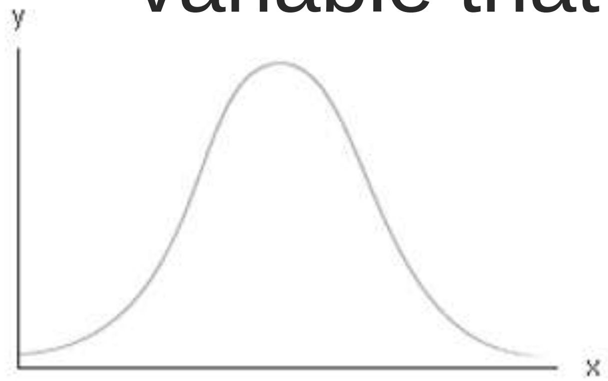
Nominal

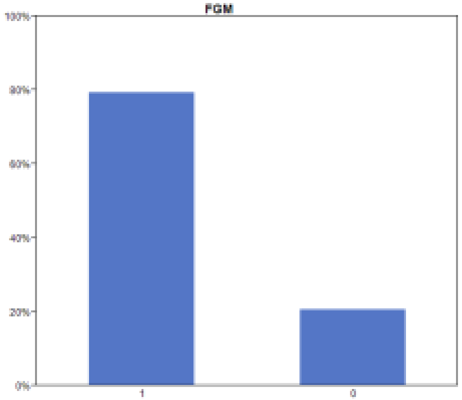
Ordinal

Why not use Regression (OLS)?

- Biggest issue is that the predicted values will take on values that have no meaning to your response

- Added mathematical inconvenience of not being able to assume normality and constant variance with the response variable that has only 2 values





Logistic Regression Model

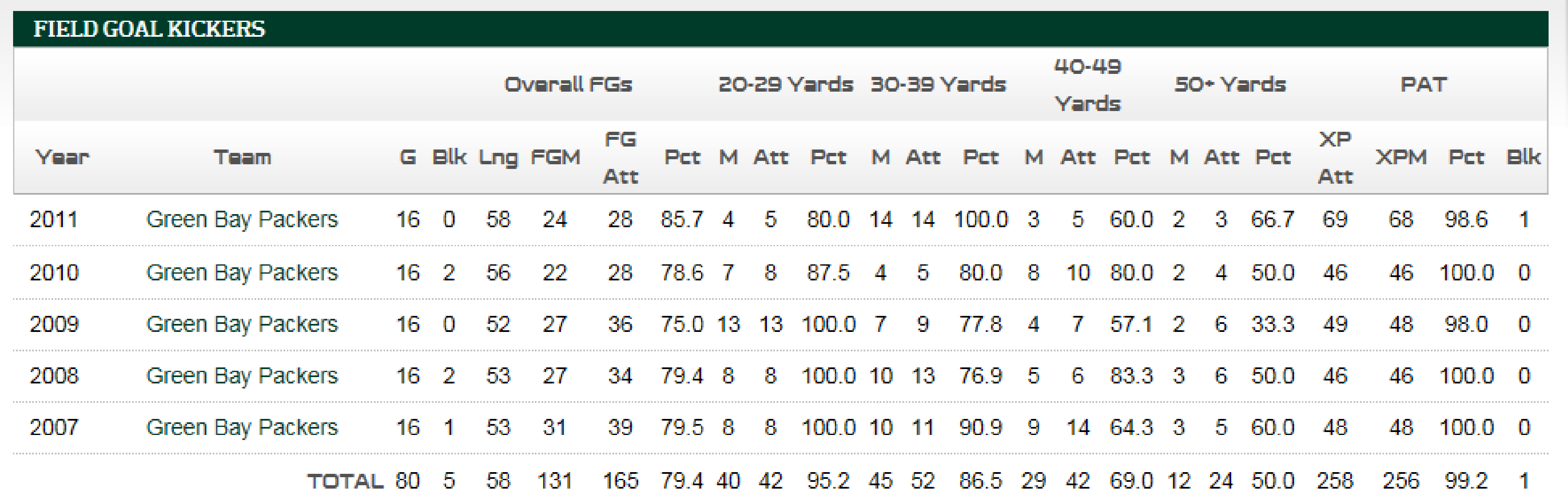
$$\text{logit}(p_i) = \beta_0 + \beta\ _1X_{1i}+\ldots+\beta\ _kx_{ki}$$

Where

- logit (p_i)=logit of the probability of the event
- $\beta\ _0$= intercept of the regression equation
- $\beta\ _k$ = parameter estimate of the kth predictor variable

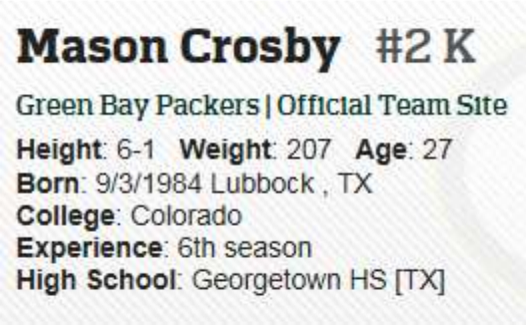
$$\text{logit}(p_i) = \log(p_i\ /\ (1-p_i))$$

Mason Crosby’s Career Field Goal Statistics



FIELD GOAL KICKERS

		Overall FGs						20-29 Yards			30-39 Yards			40-49 Yards			50+ Yards			PAT			
Year	Team	G	Blk	Lng	FGM	FG Att	Pct	M	Att	Pct	M	Att	Pct	M	Att	Pct	M	Att	Pct	XP Att	XPM	Pct	Blk
2011	Green Bay Packers	16	0	58	24	28	85.7	4	5	80.0	14	14	100.0	3	5	60.0	2	3	66.7	69	68	98.6	1
2010	Green Bay Packers	16	2	56	22	28	78.6	7	8	87.5	4	5	80.0	8	10	80.0	2	4	50.0	46	46	100.0	0
2009	Green Bay Packers	16	0	52	27	36	75.0	13	13	100.0	7	9	77.8	4	7	57.1	2	6	33.3	49	48	98.0	0
2008	Green Bay Packers	16	2	53	27	34	79.4	8	8	100.0	10	13	76.9	5	6	83.3	3	6	50.0	46	46	100.0	0
2007	Green Bay Packers	16	1	53	31	39	79.5	8	8	100.0	10	11	90.9	9	14	64.3	3	5	60.0	48	48	100.0	0
	TOTAL	80	5	58	131	165	79.4	40	42	95.2	45	52	86.5	29	42	69.0	12	24	50.0	258	256	99.2	1

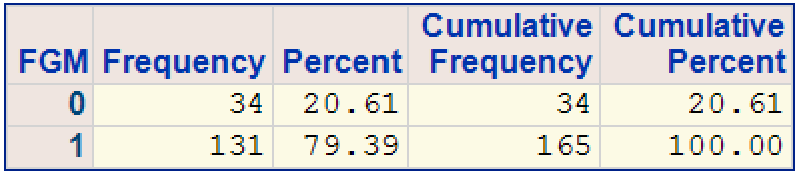


Mason Crosby #2 K

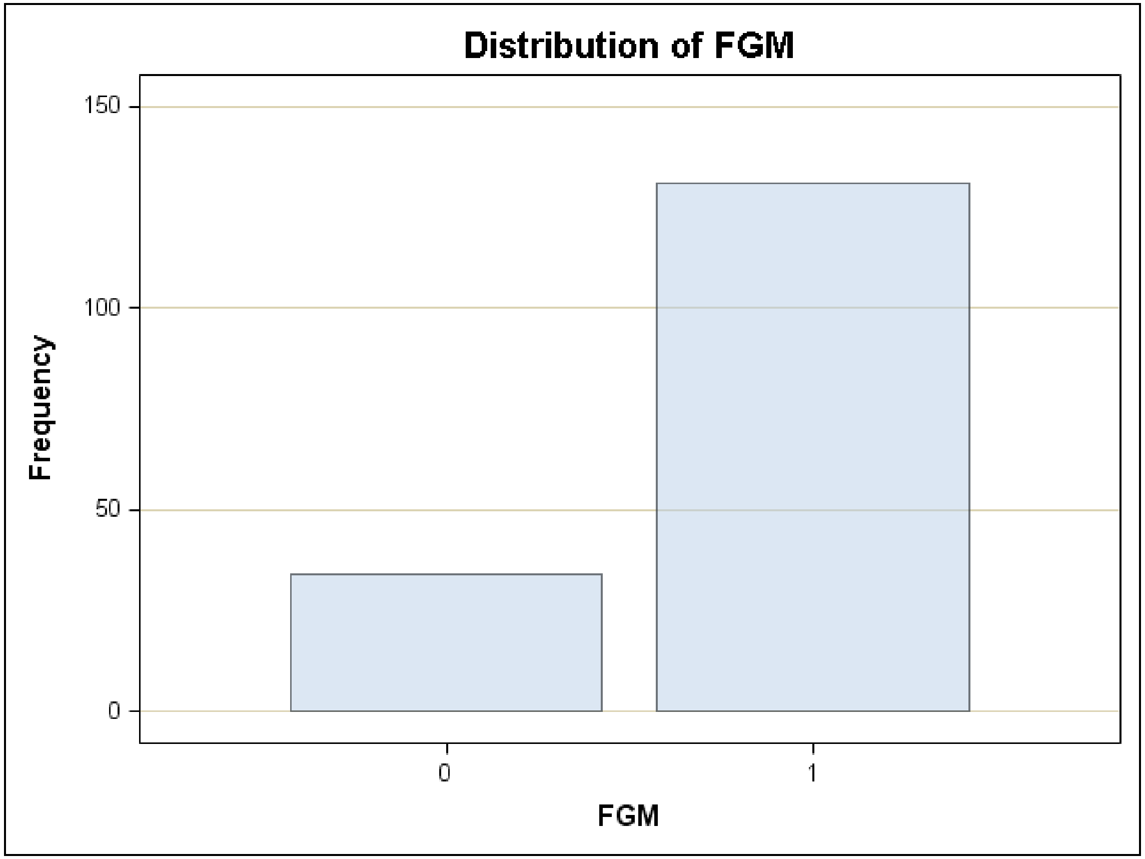
Green Bay Packers | Official Team Site

Height: 6-1 **Weight**: 207 **Age**: 27
Born: 9/3/1984 Lubbock , TX
College: Colorado
Experience: 6th season
High School: Georgetown HS [TX]

Mason Crosby’s Career Field Goal Statistics

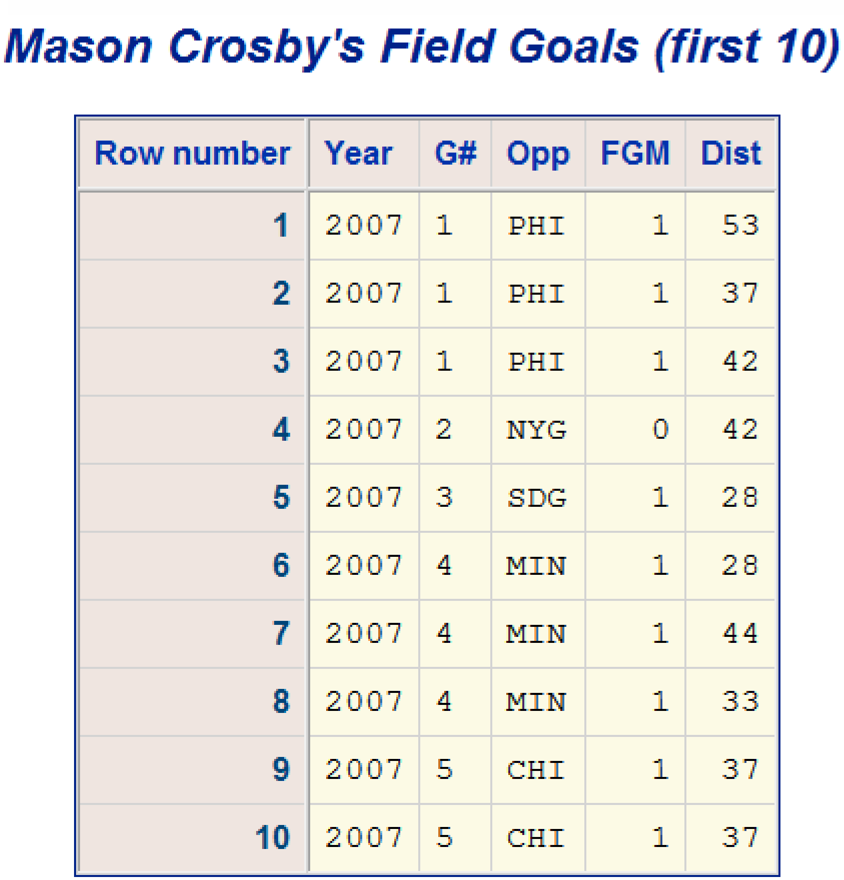


FGM	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	34	20.61	34	20.61
1	131	79.39	165	100.00



What might determine a successful field goal?

PROC LOGISTIC Data for Simple Model Continuous Predictor

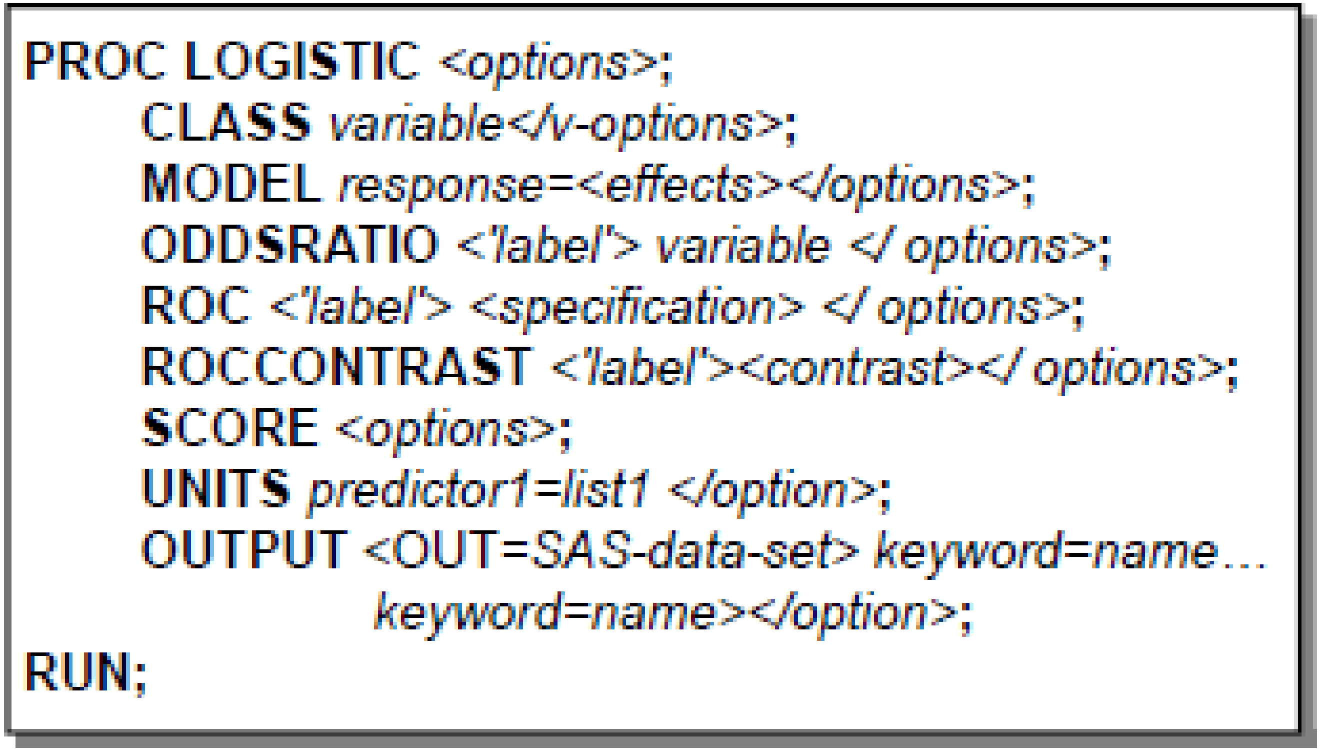


Mason Crosby's Field Goals (first 10)

Row number	Year	G#	Opp	FGM	Dist
1	2007	1	PHI	1	53
2	2007	1	PHI	1	37
3	2007	1	PHI	1	42
4	2007	2	NYG	0	42
5	2007	3	SDG	1	28
6	2007	4	MIN	1	28
7	2007	4	MIN	1	44
8	2007	4	MIN	1	33
9	2007	5	CHI	1	37
10	2007	5	CHI	1	37

Y = FGM (Field Goals Made)
X = Dist (Distance)

PROC LOGISTIC syntax

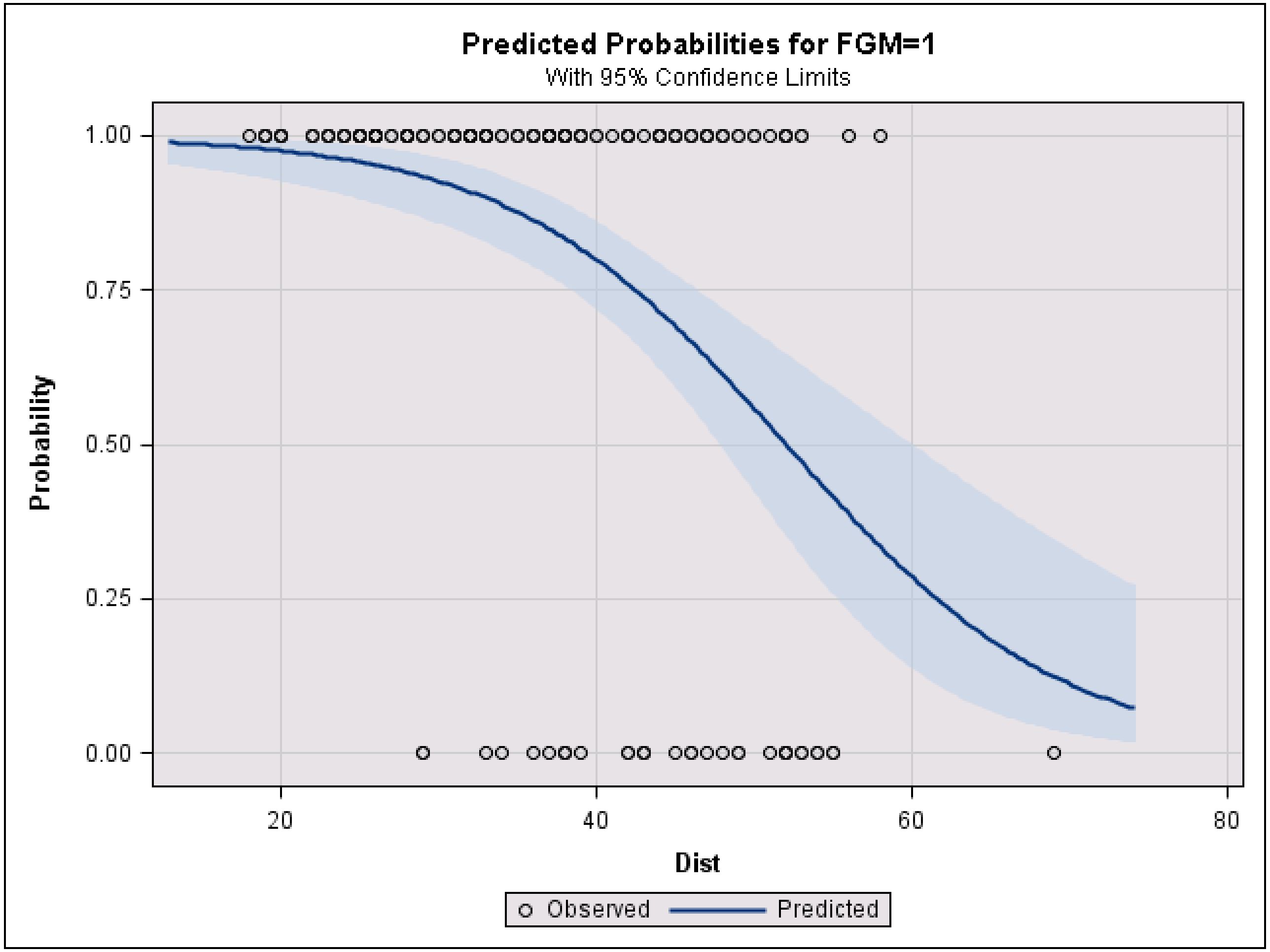


```
PROC LOGISTIC <options>;
      CLASS variable</v-options>;
      MODEL response=<effects></options>;
      ODDSRATIO <'label'> variable </ options>;
      ROC <'label'> <specification> </ options>;
      ROCCONTRAST <'label'><contrast></ options>;
      SCORE <options>;
      UNITS predictor1=list1 </option>;
      OUTPUT <OUT=SAS-data-set> keyword=name…
                  keyword=name></option>;
RUN;
```

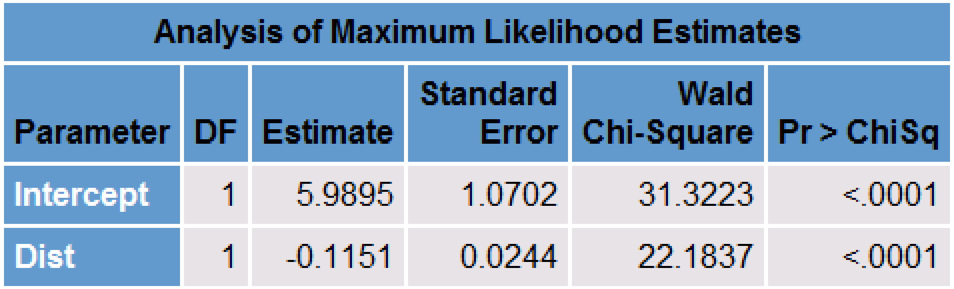
PROC LOGISTIC Code for Simple Model Continuous Predictor

```
PROC LOGISTIC DATA=WORK.Crosby_FG;
        MODEL FGM (Event = '1')=Dist/
RUN;
```

PROC LOGISTIC Output for Simple Model Continuous Predictor

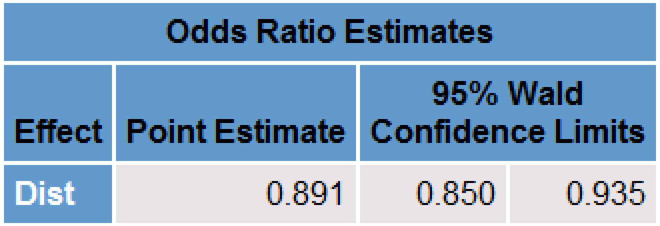


PROC LOGISTIC Output for Simple Model Continuous Predictor



Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	5.9895	1.0702	31.3223	<.0001
Dist	1	-0.1151	0.0244	22.1837	<.0001

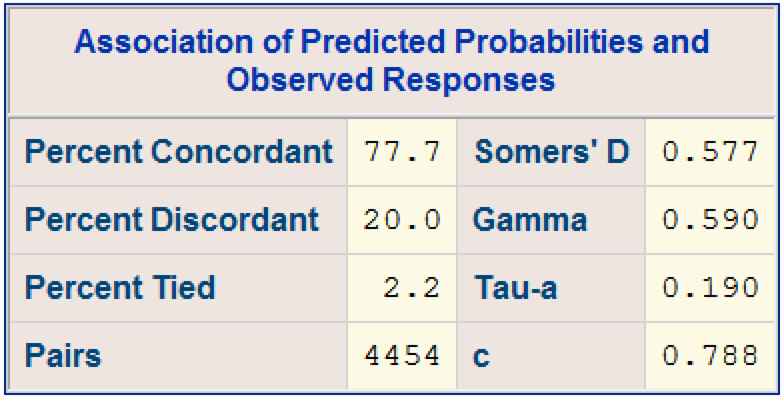


Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Dist	0.891	0.850	0.935

PROC LOGISTIC Output for Simple Model Continuous Predictor

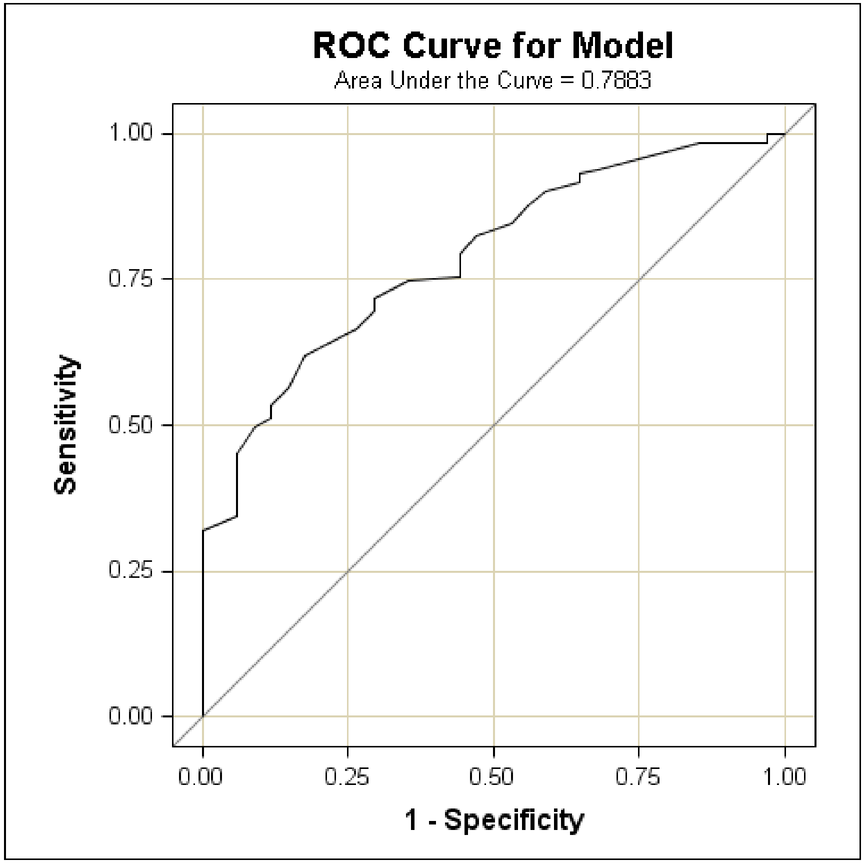
Is the model any good?



Association of Predicted Probabilities and Observed Responses

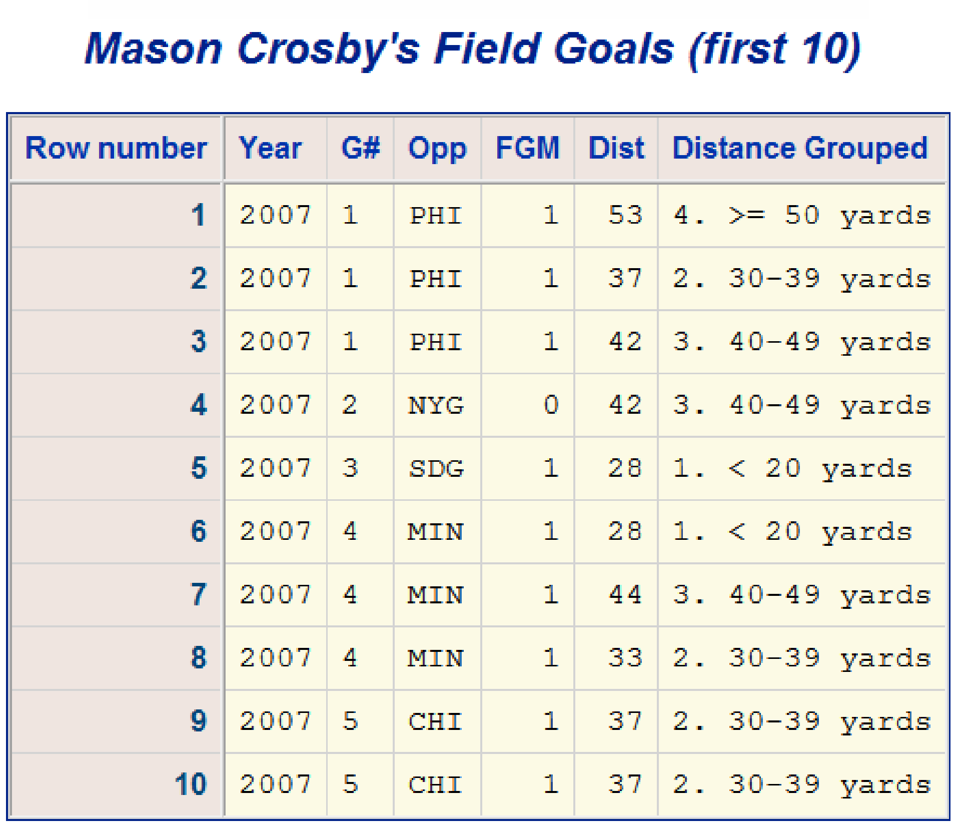
Percent Concordant	77.7	Somers' D	0.577
Percent Discordant	20.0	Gamma	0.590
Percent Tied	2.2	Tau-a	0.190
Pairs	4454	c	0.788

- Counting concordant, discordant, and tied pairs is a way to assess how well the model predicts its own data and therefore how well the model fits
- In general, you want a high percentage of concordant pairs and low percentages of discordant and tied pairs



Closer the area under the curve is to 1 the better the model, the closer to 0.5 the worse the model.

PROC LOGISTIC Data for Simple Model Categorical Predictor



Mason Crosby's Field Goals (first 10)

Row number	Year	G#	Opp	FGM	Dist	Distance Grouped
1	2007	1	PHI	1	53	4. >= 50 yards
2	2007	1	PHI	1	37	2. 30-39 yards
3	2007	1	PHI	1	42	3. 40-49 yards
4	2007	2	NYG	0	42	3. 40-49 yards
5	2007	3	SDG	1	28	1. < 20 yards
6	2007	4	MIN	1	28	1. < 20 yards
7	2007	4	MIN	1	44	3. 40-49 yards
8	2007	4	MIN	1	33	2. 30-39 yards
9	2007	5	CHI	1	37	2. 30-39 yards
10	2007	5	CHI	1	37	2. 30-39 yards

Y = FGM (Field Goals Made)
X = Dist_grp (Distance Grouped)

PROC LOGISTIC Code for Simple Model – Categorical Predictor

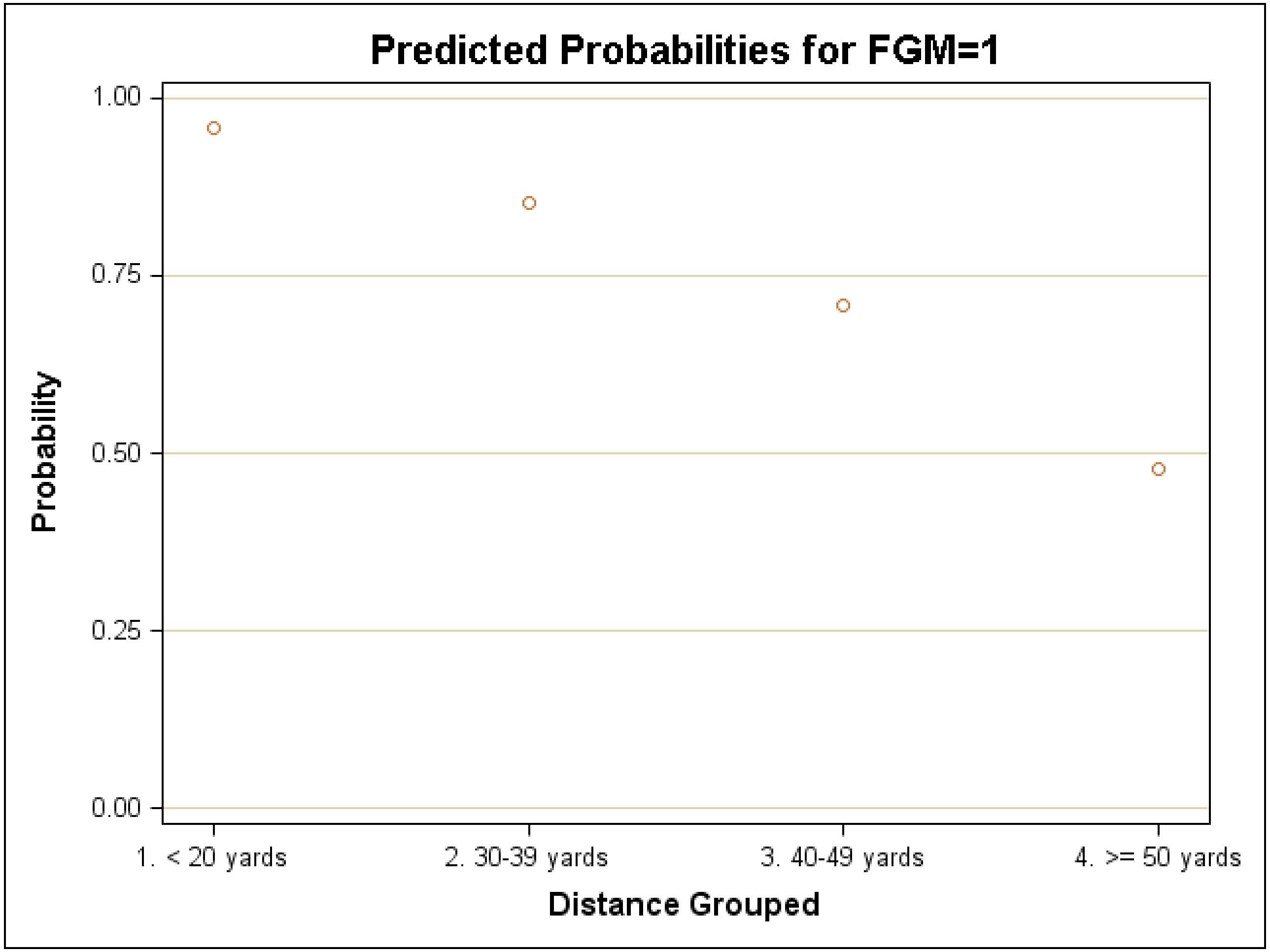
Create Categorical Variable

```
(CASE
          WHEN t1.Dist <= 29 THEN ‘1. < 29 yards’
          WHEN t1.Dist >= 30 AND t1.Dist <= 39 THEN ‘2. 30-39 yards’
          WHEN t1.Dist >= 40 AND t1.Dist <= 49 THEN ‘3. 40-49 yards’
          WHEN t1.Dist >= 50 THEN ‘4. >= 50 yards’
          ELSE t1.Dist
        END)
        LABEL="Distance Grouped" AS Dist_Grp
```

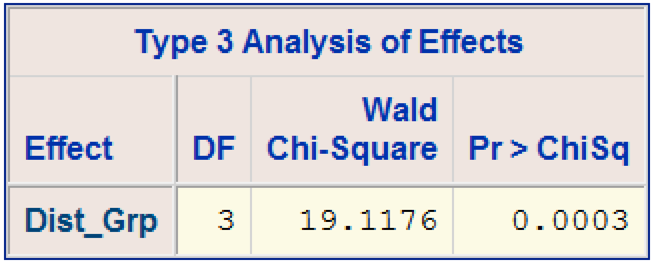
PROC LOGISTIC Code for Simple Model Categorical Predictor

```
PROC LOGISTIC DATA=WORK.Crosby_FG;
          CLASS Dist_Grp(PARAM=EFFECT);
          MODEL FGM (Event = '1')=Dist_Grp;
RUN;
```

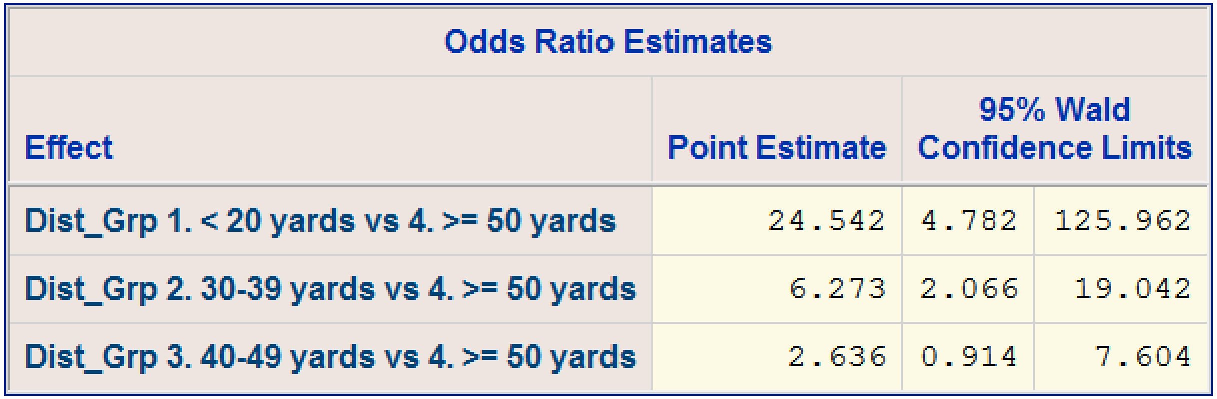
PROC LOGISTIC Output for Simple Model Categorical Predictor



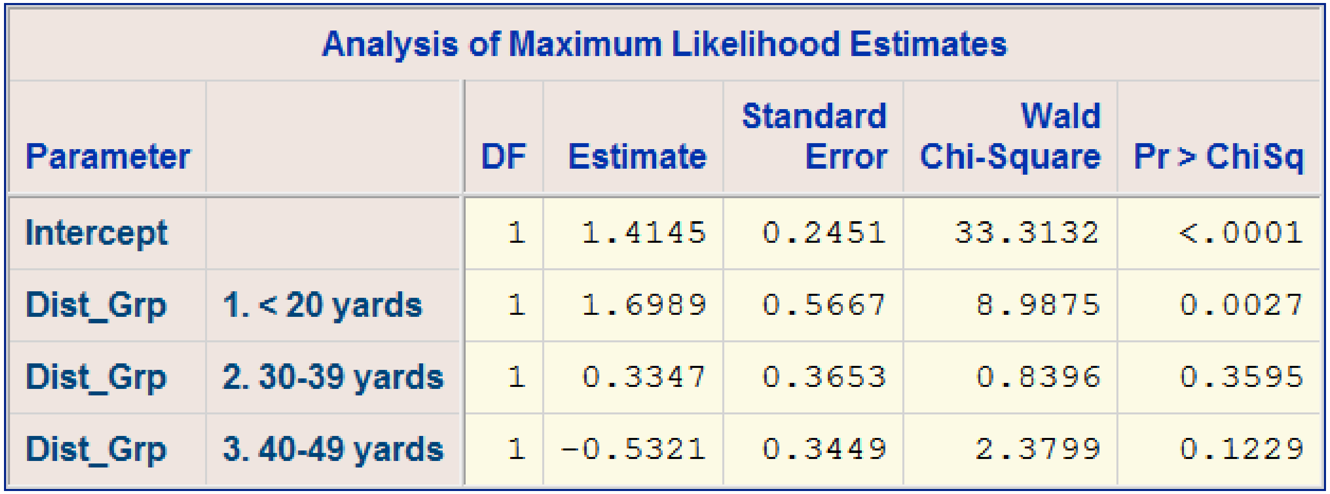
PROC LOGISTIC Code for Simple Model Categorical Predictor



Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Dist_Grp	3	19.1176	0.0003



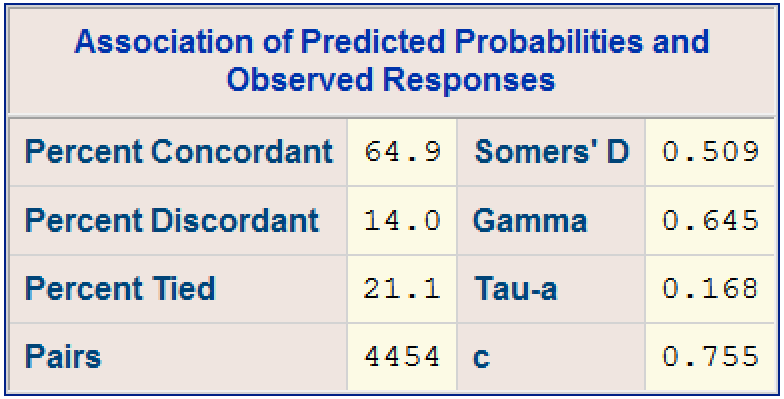
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Dist_Grp 1. < 20 yards vs 4. >= 50 yards	24.542	4.782	125.962
Dist_Grp 2. 30-39 yards vs 4. >= 50 yards	6.273	2.066	19.042
Dist_Grp 3. 40-49 yards vs 4. >= 50 yards	2.636	0.914	7.604



Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.4145	0.2451	33.3132	<.0001
Dist_Grp	1. < 20 yards	1	1.6989	0.5667	8.9875	0.0027
Dist_Grp	2. 30-39 yards	1	0.3347	0.3653	0.8396	0.3595
Dist_Grp	3. 40-49 yards	1	-0.5321	0.3449	2.3799	0.1229

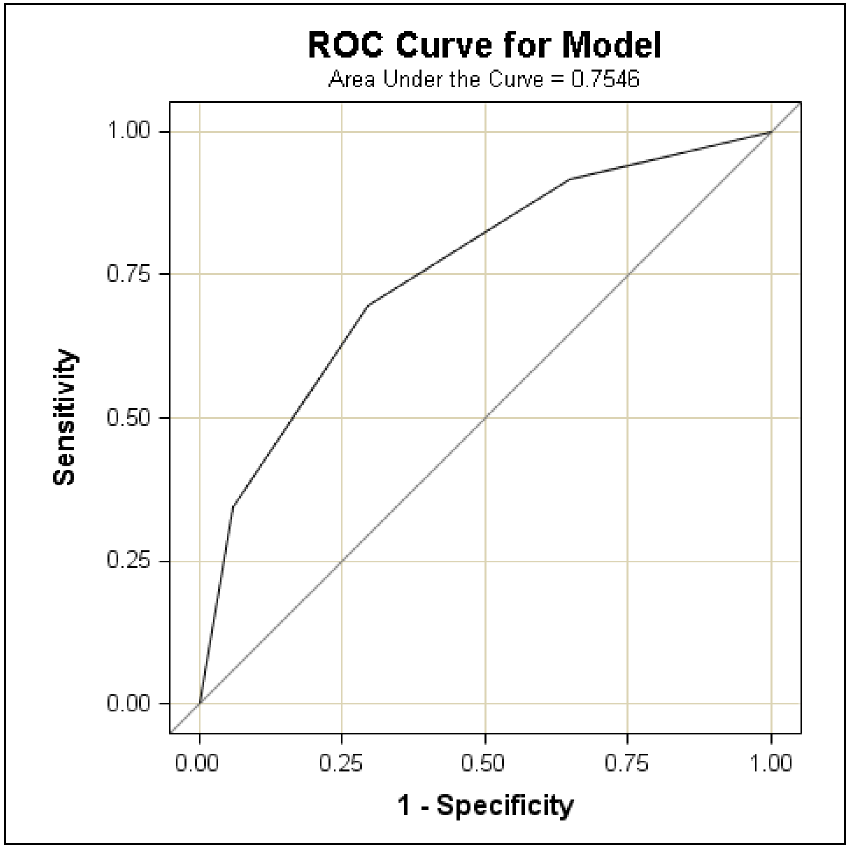
PROC LOGISTIC Output for Simple Model Categorical Predictor

Is the model any good?

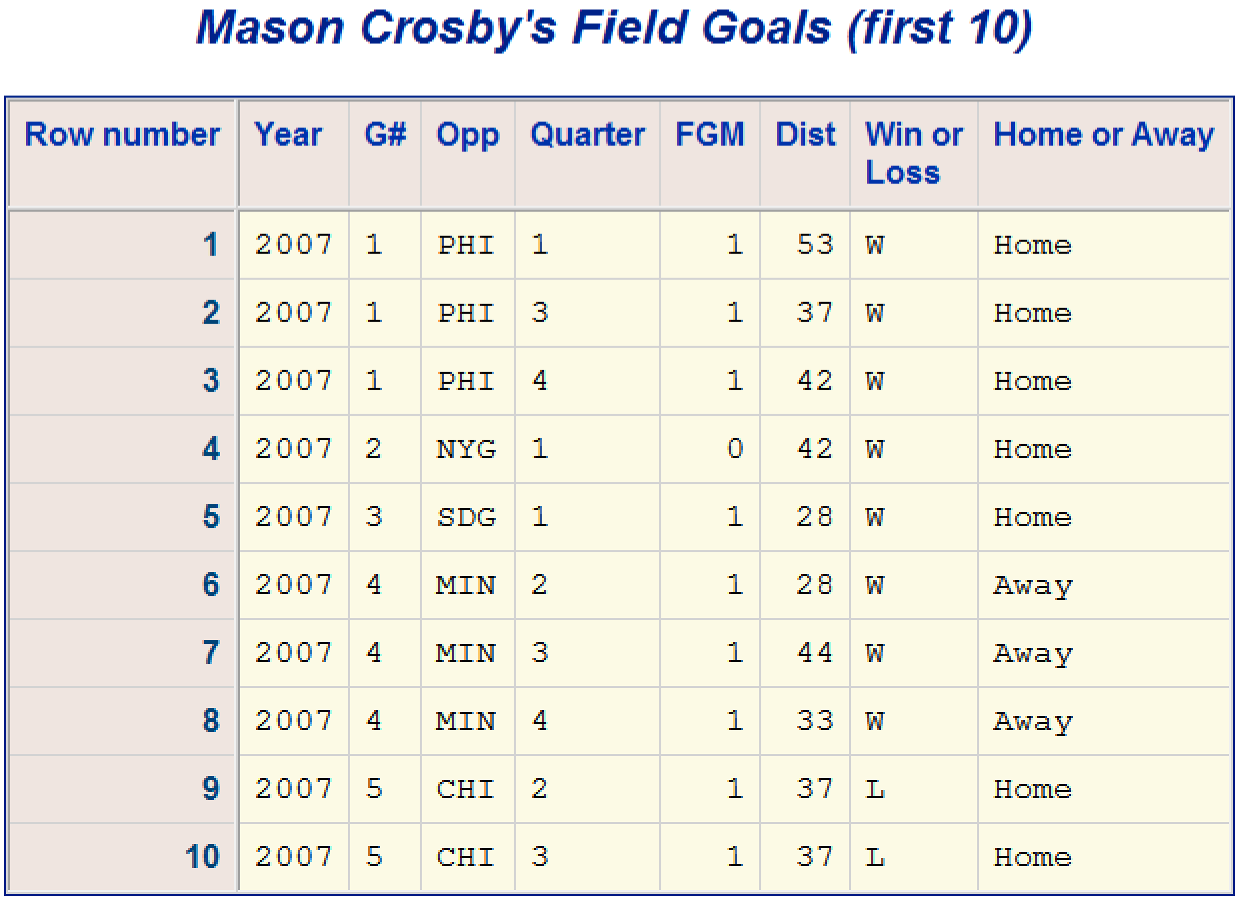


Association of Predicted Probabilities and Observed Responses			
Percent Concordant	64.9	Somers' D	0.509
Percent Discordant	14.0	Gamma	0.645
Percent Tied	21.1	Tau-a	0.168
Pairs	4454	c	0.755

Better or worse than the Continuous Model?



PROC LOGISTIC Data for Multiple Model



Mason Crosby's Field Goals (first 10)

Row number	Year	G#	Opp	Quarter	FGM	Dist	Win or Loss	Home or Away
1	2007	1	PHI	1	1	53	W	Home
2	2007	1	PHI	3	1	37	W	Home
3	2007	1	PHI	4	1	42	W	Home
4	2007	2	NYG	1	0	42	W	Home
5	2007	3	SDG	1	1	28	W	Home
6	2007	4	MIN	2	1	28	W	Away
7	2007	4	MIN	3	1	44	W	Away
8	2007	4	MIN	4	1	33	W	Away
9	2007	5	CHI	2	1	37	L	Home
10	2007	5	CHI	3	1	37	L	Home

Y = FGM (Field Goals Made)

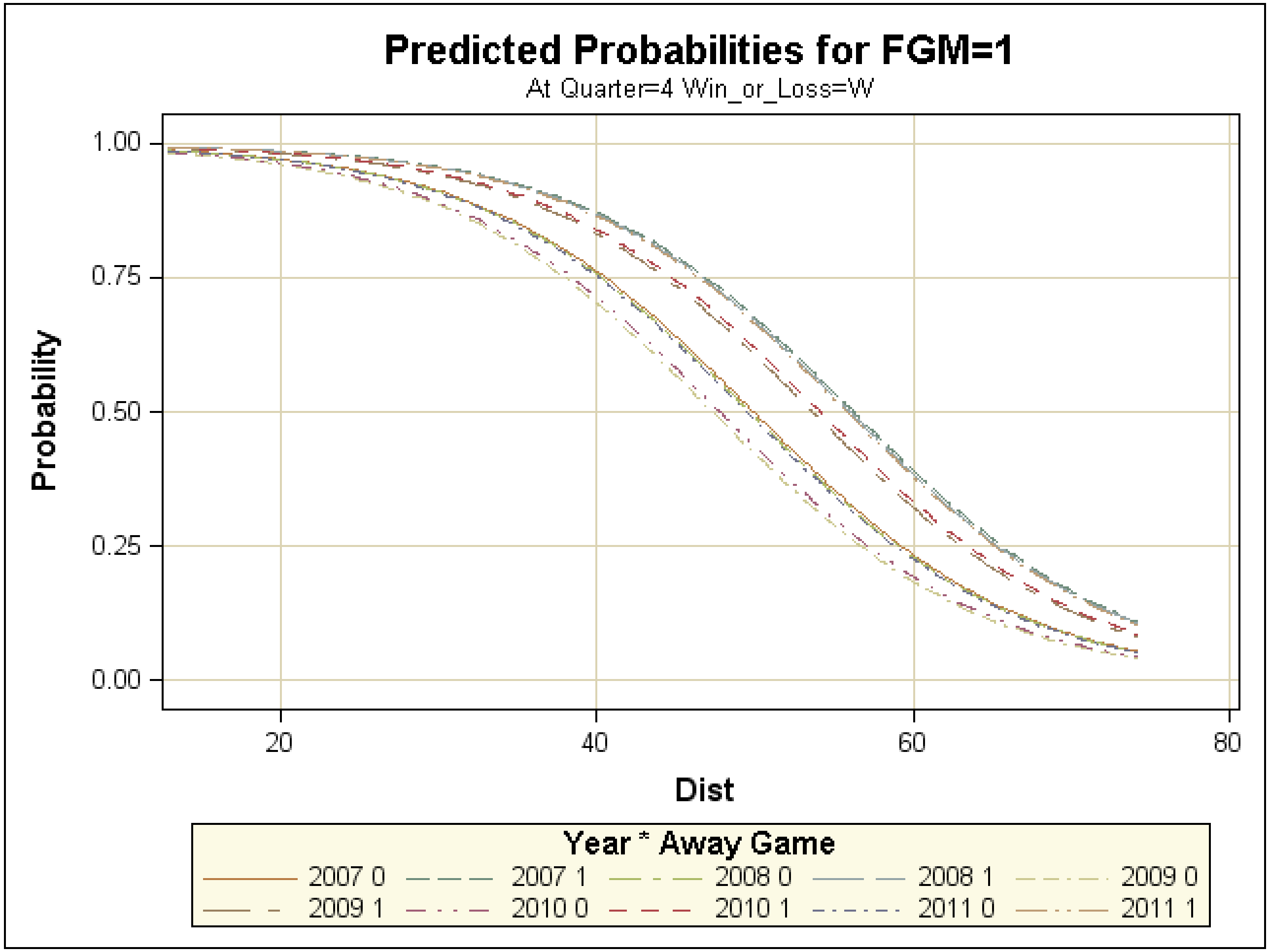
X = Dist (Distance)

Year, Quarter, Win or Loss, Home or Away

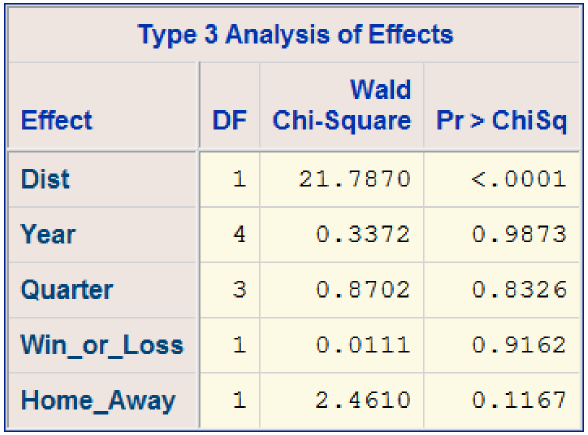
PROC LOGISTIC Code for Multiple Model

```
PROC LOGISTIC DATA=WORK.Crosby_FG;

CLASS Year Away_Game  Quarter Win_or_Loss;

MODEL FGM (Event = '1')=Dist Year Away_Game Quarter
Win_or_Loss ;

RUN;
```

PROC LOGISTIC Output for Multiple Model

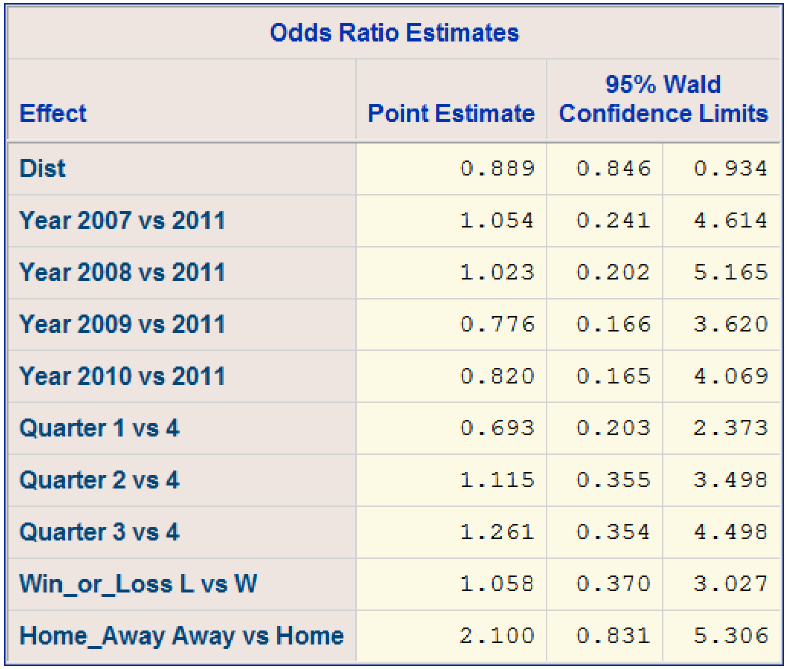


PROC LOGISTIC Code for Multiple Model



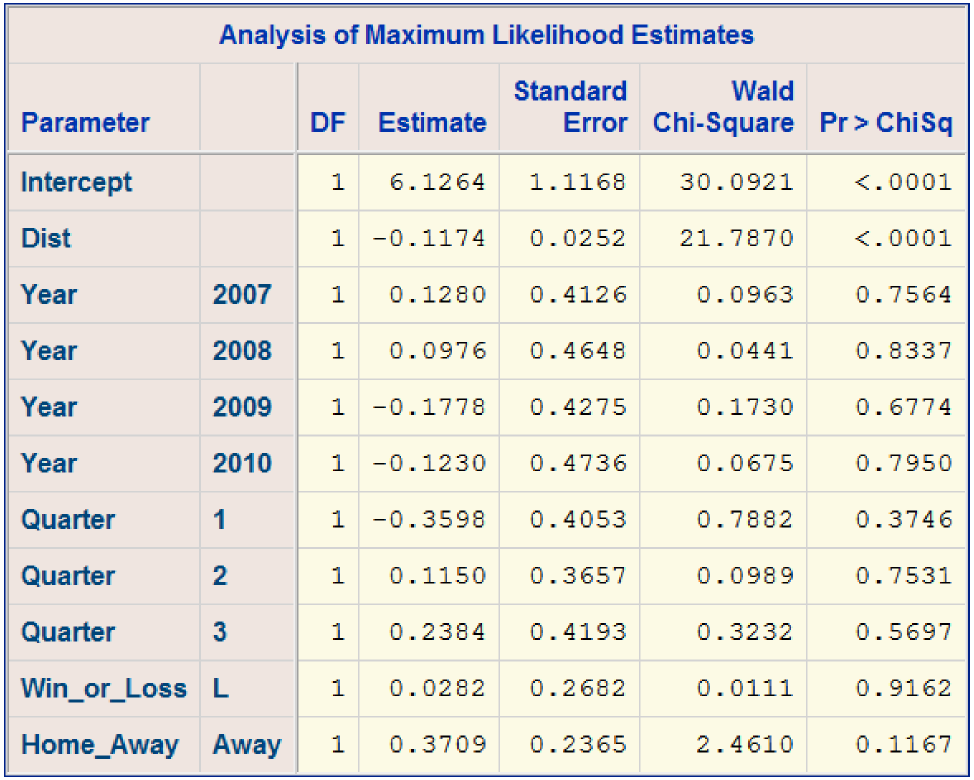
Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Dist	1	21.7870	<.0001
Year	4	0.3372	0.9873
Quarter	3	0.8702	0.8326
Win_or_Loss	1	0.0111	0.9162
Home_Away	1	2.4610	0.1167



Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Dist	0.889	0.846	0.934
Year 2007 vs 2011	1.054	0.241	4.614
Year 2008 vs 2011	1.023	0.202	5.165
Year 2009 vs 2011	0.776	0.166	3.620
Year 2010 vs 2011	0.820	0.165	4.069
Quarter 1 vs 4	0.693	0.203	2.373
Quarter 2 vs 4	1.115	0.355	3.498
Quarter 3 vs 4	1.261	0.354	4.498
Win_or_Loss L vs W	1.058	0.370	3.027
Home_Away Away vs Home	2.100	0.831	5.306

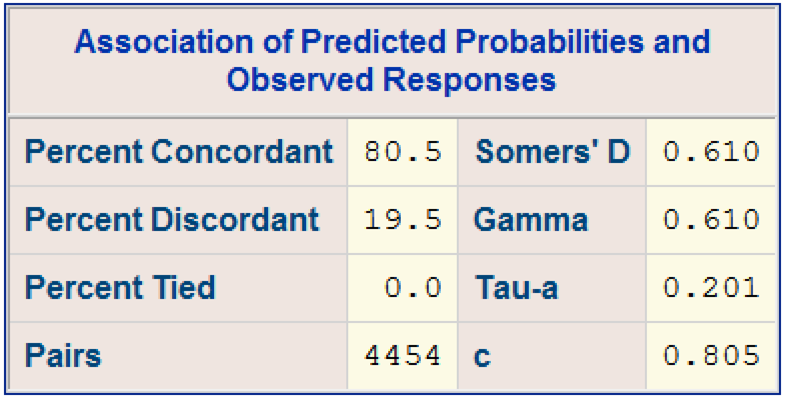


Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	6.1264	1.1168	30.0921	<.0001
Dist		1	-0.1174	0.0252	21.7870	<.0001
Year	2007	1	0.1280	0.4126	0.0963	0.7564
Year	2008	1	0.0976	0.4648	0.0441	0.8337
Year	2009	1	-0.1778	0.4275	0.1730	0.6774
Year	2010	1	-0.1230	0.4736	0.0675	0.7950
Quarter	1	1	-0.3598	0.4053	0.7882	0.3746
Quarter	2	1	0.1150	0.3657	0.0989	0.7531
Quarter	3	1	0.2384	0.4193	0.3232	0.5697
Win_or_Loss	L	1	0.0282	0.2682	0.0111	0.9162
Home_Away	Away	1	0.3709	0.2365	2.4610	0.1167

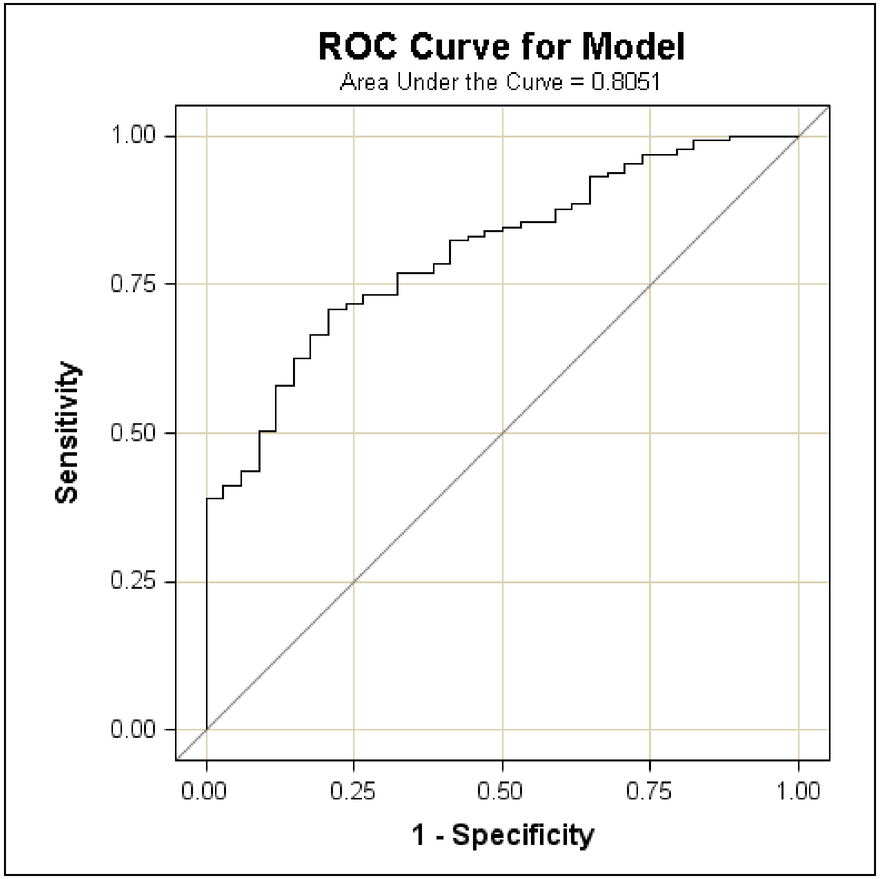
PROC LOGISTIC Output for Multiple Model

Is the model any good?



Association of Predicted Probabilities and Observed Responses

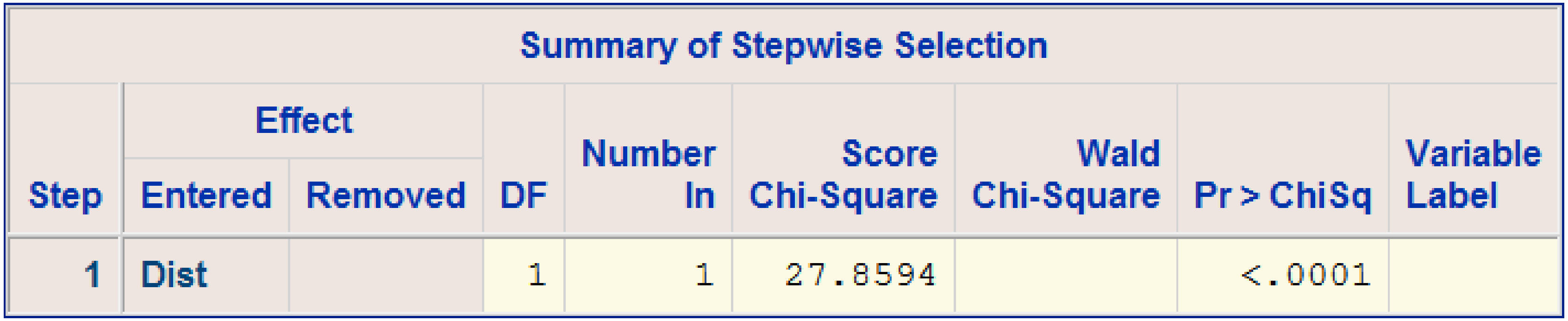
Percent Concordant	80.5	Somers' D	0.610
Percent Discordant	19.5	Gamma	0.610
Percent Tied	0.0	Tau-a	0.201
Pairs	4454	c	0.805



Better or worse than the Simple Models?

Stepwise Options

- Forward
- Backward
- Stepwise



Summary of Stepwise Selection

	Effect			Number	Score	Wald		Variable
Step	Entered	Removed	DF	In	Chi-Square	Chi-Square	Pr > ChiSq	Label
1	Dist		1	1	27.8594		<.0001	

Challenges

- Missing Value
- Errors and Outliers
- Massive Data size
- Operational vs. observational

It’s Good!

Resources

Public SAS Courses

- Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression
- Predictive Modeling Using Logistic Regression
- Categorical Data Analysis Using Logistic Regression

Books

- *Logistic Regression Using SAS Theory and Application, Second Edition* by Paul D Allison

Online Tutorials

- Logistic Regression in SAS Enterprise Guide Example 1
- Logistic Regression in SAS Enterprise Guide Example 2

It's where you'll find the latest information on:

- New training courses and services
- Special offers and discounts
- The latest course schedules
- New training locations
- Events and conferences
- SAS certification news
- And, much more.

Everything you need – in one place.
Visit and bookmark it today.

THE POWER TO KNOW

Thank you for using SAS!

www.sas.com