CATEGORICAL DATA ANALYSIS

Institute for Advanced Analytics

MSA Class of 2020

DESCRIBING CATEGORICAL DATA

Examining Categorical Variables

- By examining the distributions of categorical variables, you can do the following:
  1. Determine the frequencies of data values
  2. Recognize possible associations among variables

Categorical Variables Association

- An association exists between two categorical variables if the distribution of one variable changes when the level (or value) of the other variable changes.
- If there is no association, the distribution of the first variable is the same regardless of the level of the other variable.

No Association

	72%	28%
	72%	28%

Is your manager’s mood associated with the weather?

Association

		
	82%	18%
	60%	40%

Is your manager’s mood associated with the weather?

Frequency Tables

- A frequency table shows the number of observations that occur in certain categories or intervals. A one-way frequency table examines one variable.

Income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High	155	36	155	36
Low	132	31	287	67
Medium	144	33	431	100

Crosstabulation Tables

- A *crosstabulation* table shows the number of observations for each combination of the row and column variables.

	column 1	column 2	…	column c
row 1	cell_{11}	cell_{12}	…	cell_{1c}
row 2	cell_{21}	cell_{22}	…	cell_{2c}
…	…	…	…	…
row r	cell_{r1}	cell_{r2}	…	cell_{rc}

Ames Housing– Bonus Eligible Sale

- Realtors in Ames, Iowa receive the standard 3% commission on homes sales. One particular realty company offers an additional bonus for homes that sell for more than $175,000. Are there attributes of the home that can predict whether it will be bonus eligible?

Examining Distributions – Part 1

```
proc freq data=bootcamp.ameshousing3;
    tables Bonus Fireplaces Lot_Shape_2
           Fireplaces*Bonus Lot_Shape_2*Bonus/
           plots(only)=freqplot(scale=percent);
    format Bonus bonusfmt.;
run;
```

Examining Distributions – Part 1

The FREQ Procedure

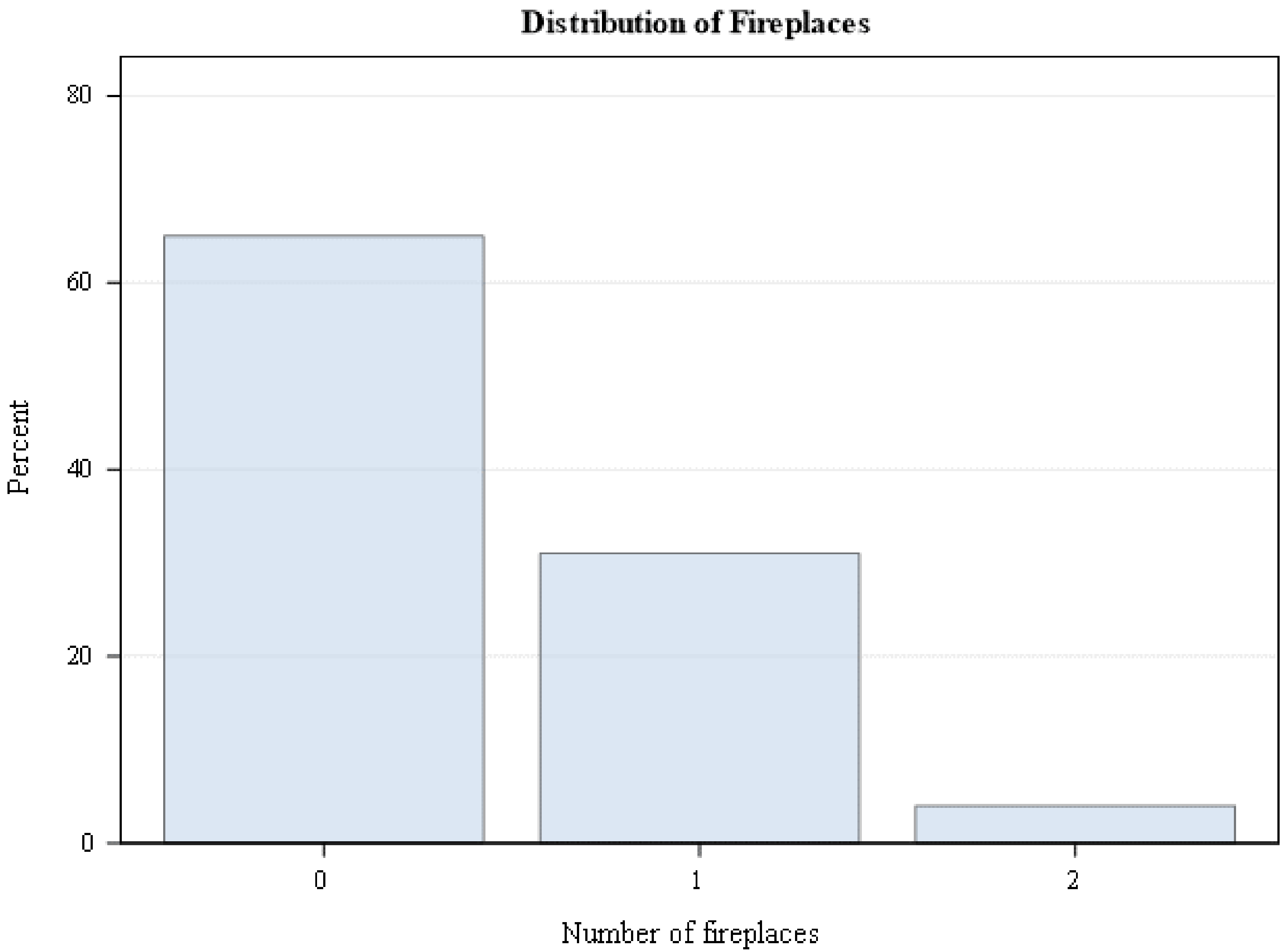
Sale Price > $175,000				
Bonus	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Not Bonus Eligible	255	85.00	255	85.00
Bonus Eligible	45	15.00	300	100.00

Examining Distributions – Part 1

Examining Distributions – Part 1

Number of fireplaces				
Fireplaces	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	195	65.00	195	65.00
1	93	31.00	288	96.00
2	12	4.00	300	100.00

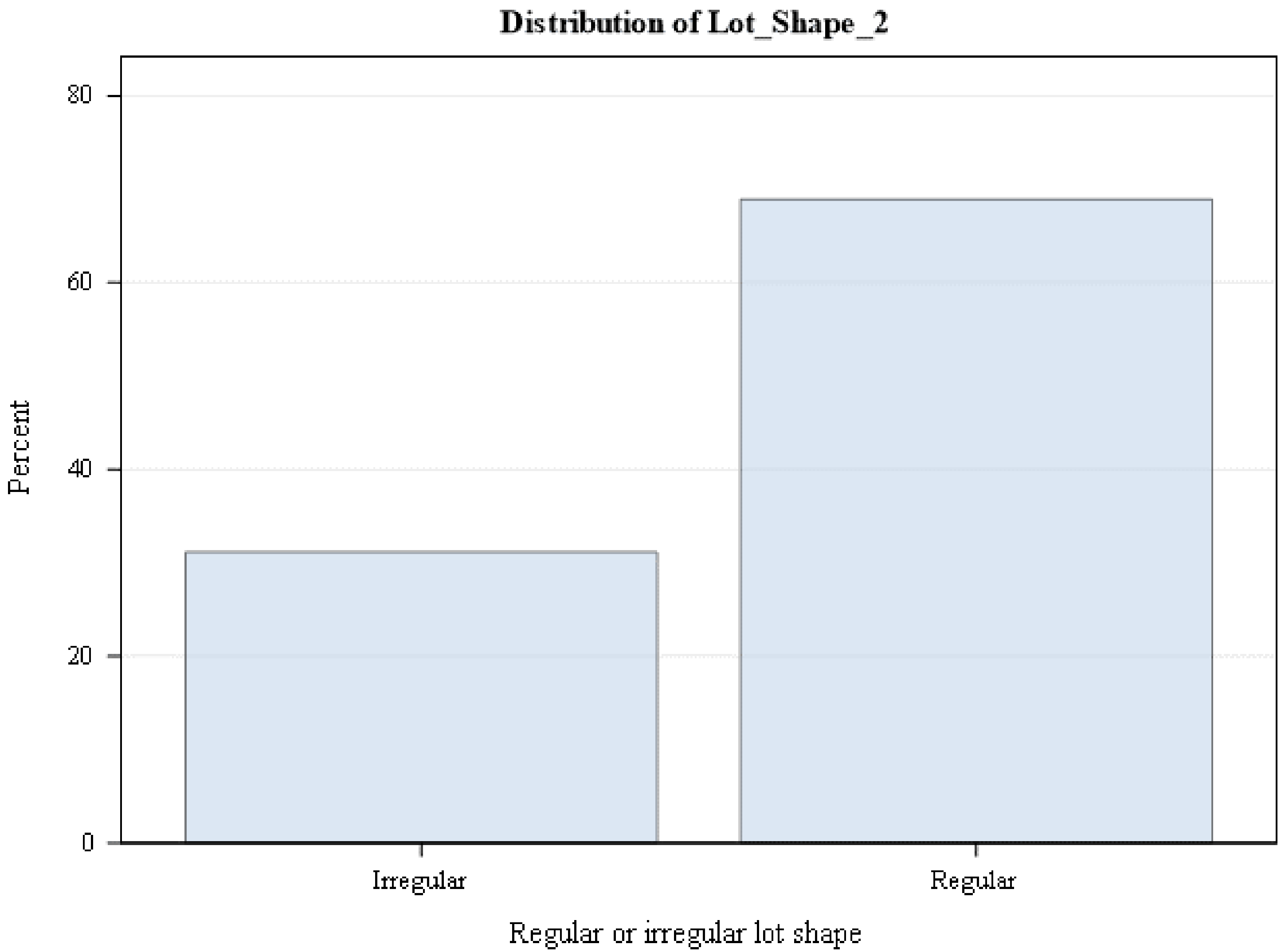
Examining Distributions – Part 1



Examining Distributions – Part 1

Regular or irregular lot shape				
Lot_Shape_2	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Irregular	93	31.10	93	31.10
Regular	206	68.90	299	100.00
Frequency Missing = 1				

Examining Distributions – Part 1

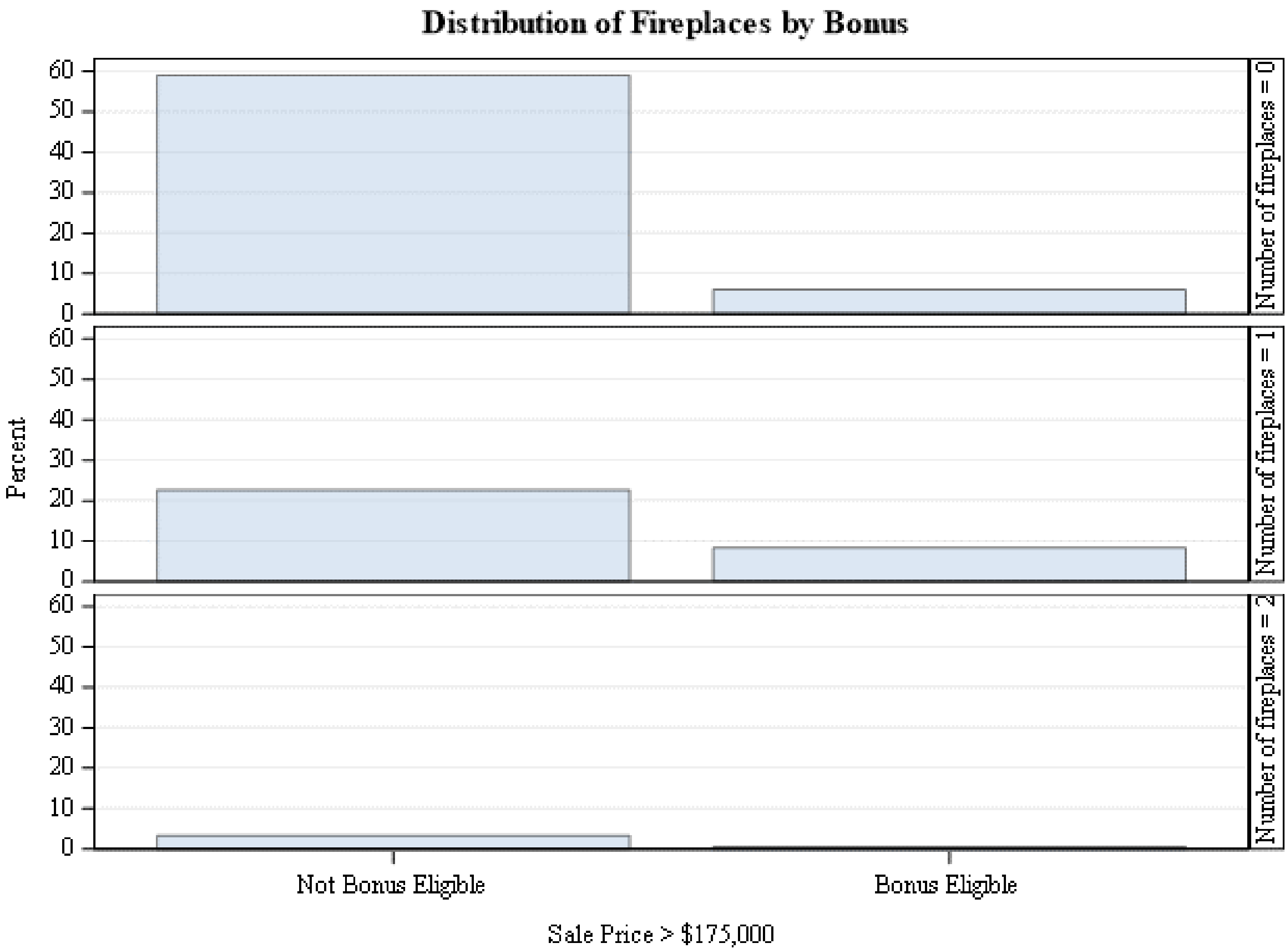


Examining Distributions – Part 1

Frequency Percent Row Pct Col Pct

Table of Fireplaces by Bonus			
Fireplaces(Number of fireplaces)	Bonus(Sale Price > $175,000)		
	Not Bonus Eligible	Bonus Eligible	Total
0	177 59.00 90.77 69.41	18 6.00 9.23 40.00	195 65.00
1	68 22.67 73.12 26.67	25 8.33 26.88 55.56	93 31.00
2	10 3.33 83.33 3.92	2 0.67 16.67 4.44	12 4.00
Total	255 85.00	45 15.00	300 100.00

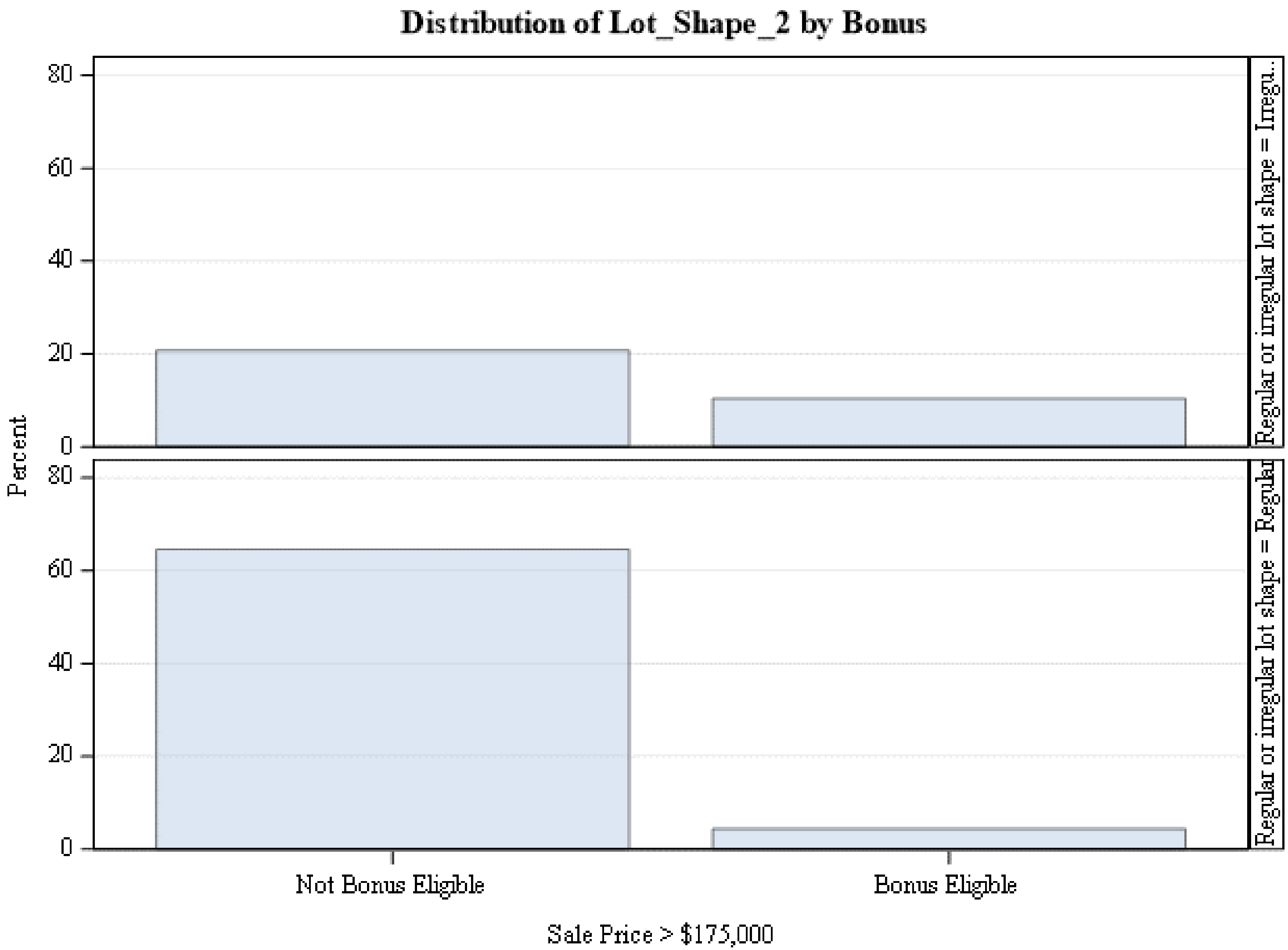
Examining Distributions – Part 1



Examining Distributions – Part 1

Table of Lot_Shape_2 by Bonus			
Lot_Shape_2(Regular or irregular lot shape)	Bonus(Sale Price > $175,000)		
	0	1	Total
Irregular	62 20.74 66.67 24.31	31 10.37 33.33 70.45	93 31.10
Regular	193 64.55 93.69 75.69	13 4.35 6.31 29.55	206 68.90
Total	255 85.28	44 14.72	299 100.00
Frequency Missing = 1			

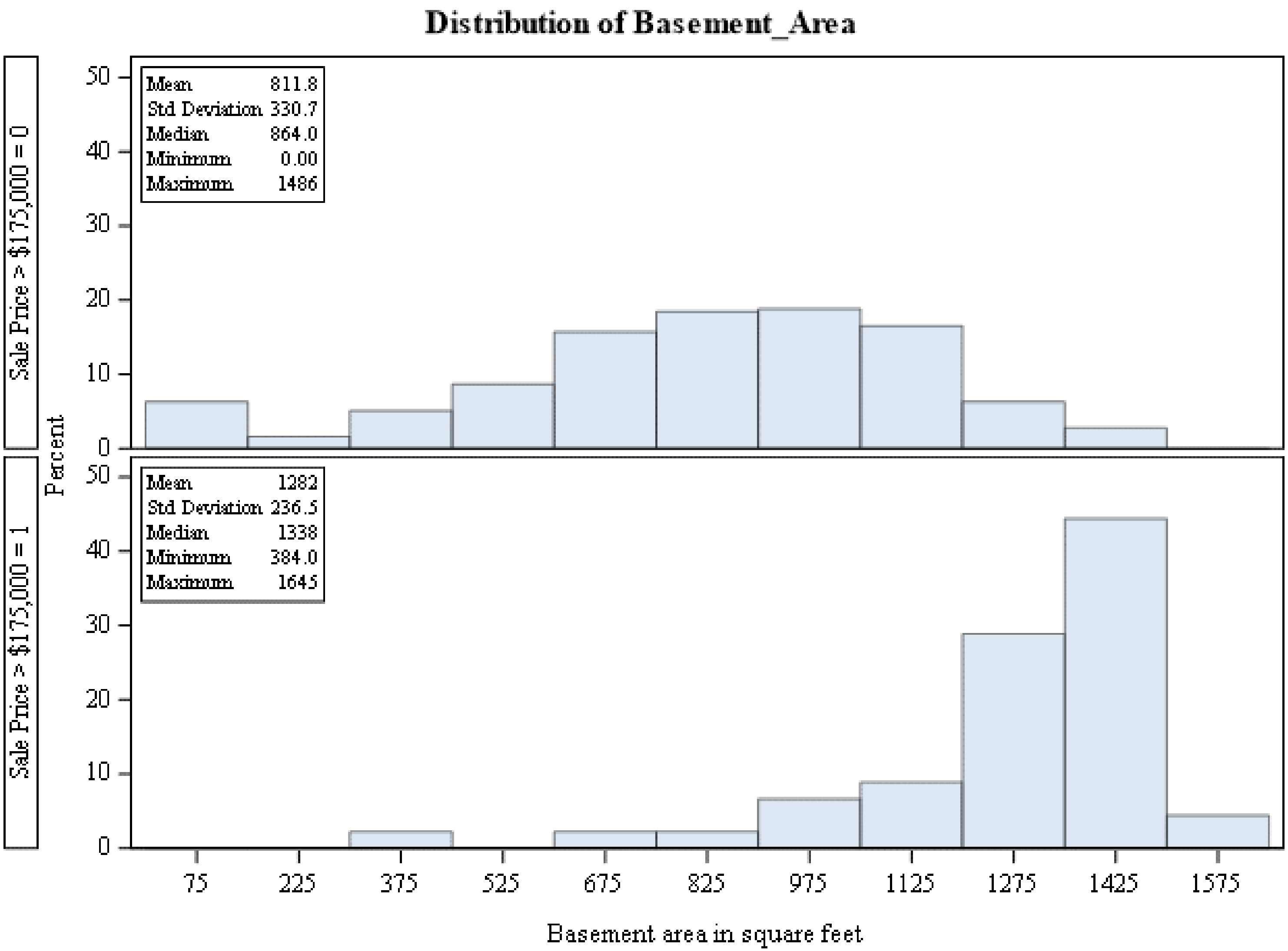
Examining Distributions – Part 1



Examining Distributions – Part 2

```
proc univariate data=bootcamp.ameshousing3 noprint;
    class Bonus;
    var Basement_Area ;
    histogram Basement_Area;
    inset mean std median min max / format=5.2 position=nw;
    format Bonus bonusfmt.;
run;
```

Examining Distributions – Part 2



?

TESTS OF ASSOCIATION

Introduction

Table of Lot_Shape_2 by Bonus			
Lot_Shape_2	Bonus		
Row Pct	Not Bonus Eligible	Bonus Eligible	Total
Irregular	**66.67%**	**33.33%**	N=93
Regular	**93.69%**	**6.31%**	N=206
Total	N=255	N=44	N=299

Tests of Association - Hypotheses

- **Null Hypothesis**
  - There is no association between **Lot_Shape_2** and **Bonus**.
  - The probability of a home sale being bonus eligible is the same regardless of lot shape.

- **Alternative Hypothesis**
  - There ***is*** an association between **Lot_Shape_2** and **Bonus**.
  - The probability of a home sale being bonus eligible is not the same for irregular and regular lot shapes.

Chi-Square Test

NO ASSOCIATION

observed frequencies=expected frequencies

ASSOCIATION

observed frequencies≠expected frequencies

The expected frequencies are calculated by the formula: (row total*column total) / sample size.

Chi-Square Tests

- Chi-square tests and the corresponding *p*-values:
  - Determine whether an association exists
  - DO NOT measure the strength of an association
  - Depend on and reflect the sample size

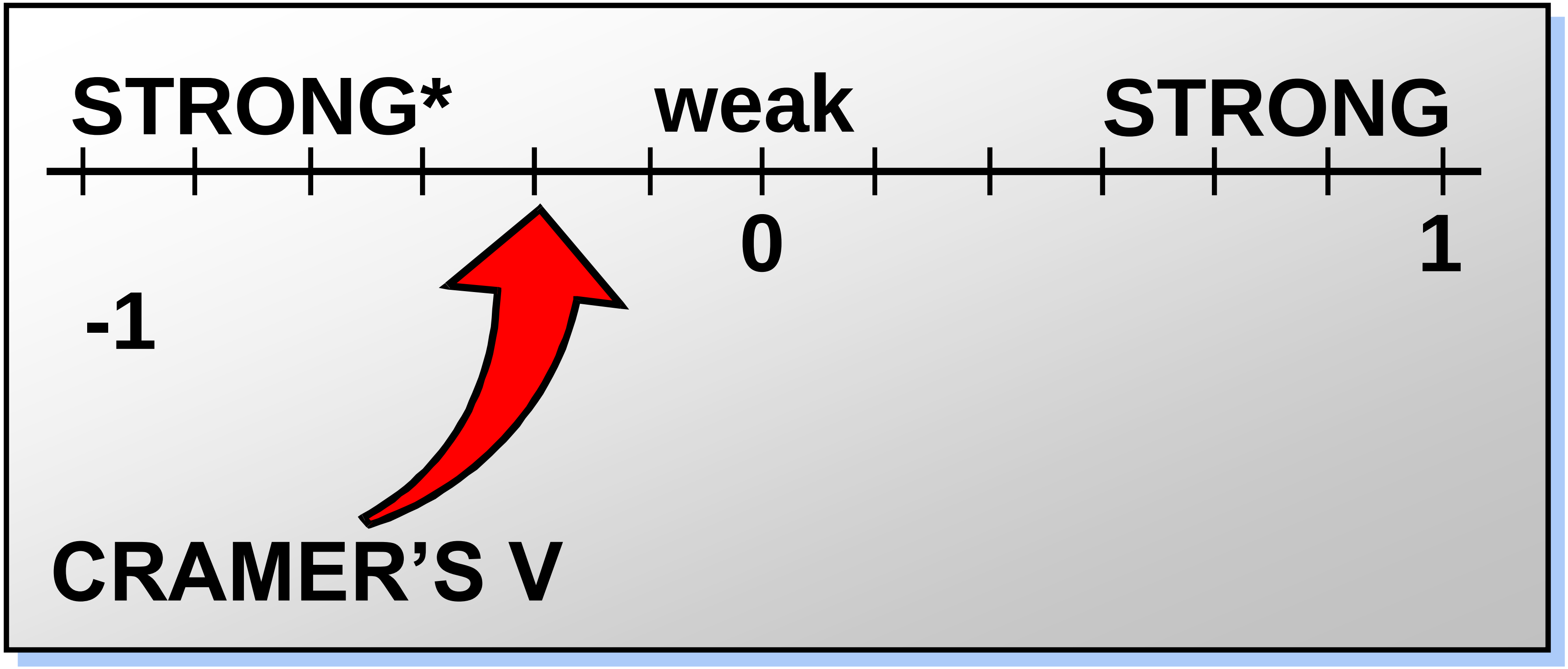
$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{\left(Obs_{i,j} - Exp_{i,j}\right)^2}{Exp_{i,j}}$$

Chi-Square Tests

- Sample size requirements:
  - 80% or more of the cells need **expected** count larger than 5.

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{\left(Obs_{i,j} - Exp_{i,j}\right)^2}{Exp_{i,j}}$$

Measures of Association



*** Cramer’s V is always nonnegative for tables larger than 2*2.**

Odds Ratios

- An *odds ratio* indicates how much more likely, with respect to odds, a certain event occurs in one group relative to its occurrence in another group.
- How do the odds of irregular lot shapes being bonus eligible compare to those of regular lot shapes?

$$\text{Odds} = \frac{p_{\text{event}}}{1 - p_{\text{event}}}$$

Probability versus Odds of an Outcome

	Yes	No	Total
Group A	60	20	**80**
Group B	90	10	**100**
Total	**150**	**30**	**180**

Probability versus Odds of an Outcome

	Yes	No	Total
Group A	60	20	**80**
Group B	90	10	**100**
Total	**150**	**30**	**180**

Probability of **YES** in **Group B** $= \dfrac{90}{100} = 0.90$

Probability versus Odds of an Outcome

	Yes	No	Total
Group A	60	20	**80**
Group B	90	10	**100**
Total	**150**	**30**	**180**

Probability of **NO** in **Group B** $= \frac{10}{100} = 0.10$

Probability versus Odds of an Outcome

	Yes	No	Total
Group A	60	20	**80**
Group B	90	10	**100**
Total	**150**	**30**	**180**

Odds of **YES** in **Group B** $= \frac{\text{Prob. of Yes}}{\text{Prob. of No}} = \frac{0.90}{0.10} = 9$

Probability versus Odds of an Outcome

	Yes	No	Total
Group A	60	20	80
Group B	90	10	100
Total	150	30	180

Odds of **YES** in **Group A** $= 3$

Odds of **YES** in **Group B** $= 9$

Odds Ratio: Group B to Group A $= \frac{9}{3} = 3$

Odds Ratio

Odds of **YES** in **Group A** $= 3$

Odds of **YES** in **Group B** $= 9$

Odds Ratio: Group B to Group A $= \frac{9}{3} = 3$

Group B observations have **3 times the odds** of having the outcome (Yes) as compared to the observations in Group A.

Chi-Square Test

```
proc freq data=bootcamp.ameshousing3;
    tables (Lot_Shape_2 Fireplaces)*Bonus
          / chisq expected cellchi2 nocol nopercent
            relrisk;
    format Bonus bonusfmt.;
    title 'Associations with Bonus';
run;
```

Chi-Square Test

Associations with Bonus
The FREQ Procedure

Frequency Expected Cell Chi-Square Row Pct

Table of Lot_Shape_2 by Bonus			
Lot_Shape_2(Regular or irregular lot shape)	Bonus(Sale Price > $175,000)		
	0	1	Total
Irregular	62 79.314 3.7797 66.67	31 13.686 21.905 33.33	93
Regular	193 175.69 1.7064 93.69	13 30.314 9.8893 6.31	206
Total	255	44	299
Frequency Missing = 1			

Chi-Square Test

Statistics for Table of Lot_Shape_2 by Bonus

Statistic	DF	Value	Prob
Chi-Square	1	37.2807	<.0001
Likelihood Ratio Chi-Square	1	34.4226	<.0001
Continuity Adj. Chi-Square	1	35.1587	<.0001
Mantel-Haenszel Chi-Square	1	37.1561	<.0001
Phi Coefficient		-0.3531	
Contingency Coefficient		0.3330	
Cramer's V		-0.3531	

Chi-Square Test

Fisher's Exact Test	
Cell (1,1) Frequency (F)	62
Left-sided Pr <= F	<.0001
Right-sided Pr >= F	1.0000
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Odds Ratio and Relative Risks			
Statistic	**Value**	**95% Confidence Limits**	
Odds Ratio	0.1347	0.0664	0.2735
Relative Risk (Column 1)	0.7116	0.6137	0.8251
Relative Risk (Column 2)	5.2821	2.9002	9.6202

Chi-Square Test

Frequency Expected Cell Chi-Square Row Pct

Table of Fireplaces by Bonus			
Fireplaces(Number of fireplaces)	Bonus(Sale Price > $175,000)		
	0	1	Total
0	177 165.75 0.7636 90.77	18 29.25 4.3269 9.23	195
1	68 79.05 1.5446 73.12	25 13.95 8.7529 26.88	93
2	10 10.2 0.0039 83.33	2 1.8 0.0222 16.67	12
Total	255	45	300

Chi-Square Test

Statistics for Table of Fireplaces by Bonus

Statistic	DF	Value	Prob
Chi-Square	2	15.4141	0.0004
Likelihood Ratio Chi-Square	2	14.4859	0.0007
Mantel-Haenszel Chi-Square	1	10.7458	0.0010
Phi Coefficient		0.2267	
Contingency Coefficient		0.2211	
Cramer's V		0.2267	

Sample Size = 300

Association among Ordinal Variables

Is

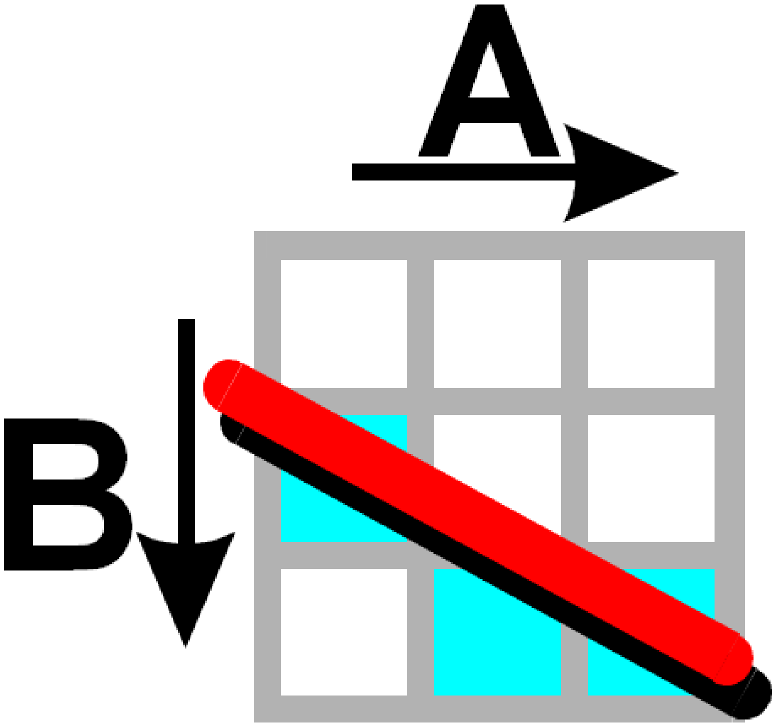
Fireplaces

associated with

Bonus

?

Mantel-Haenszel Chi-Square Test

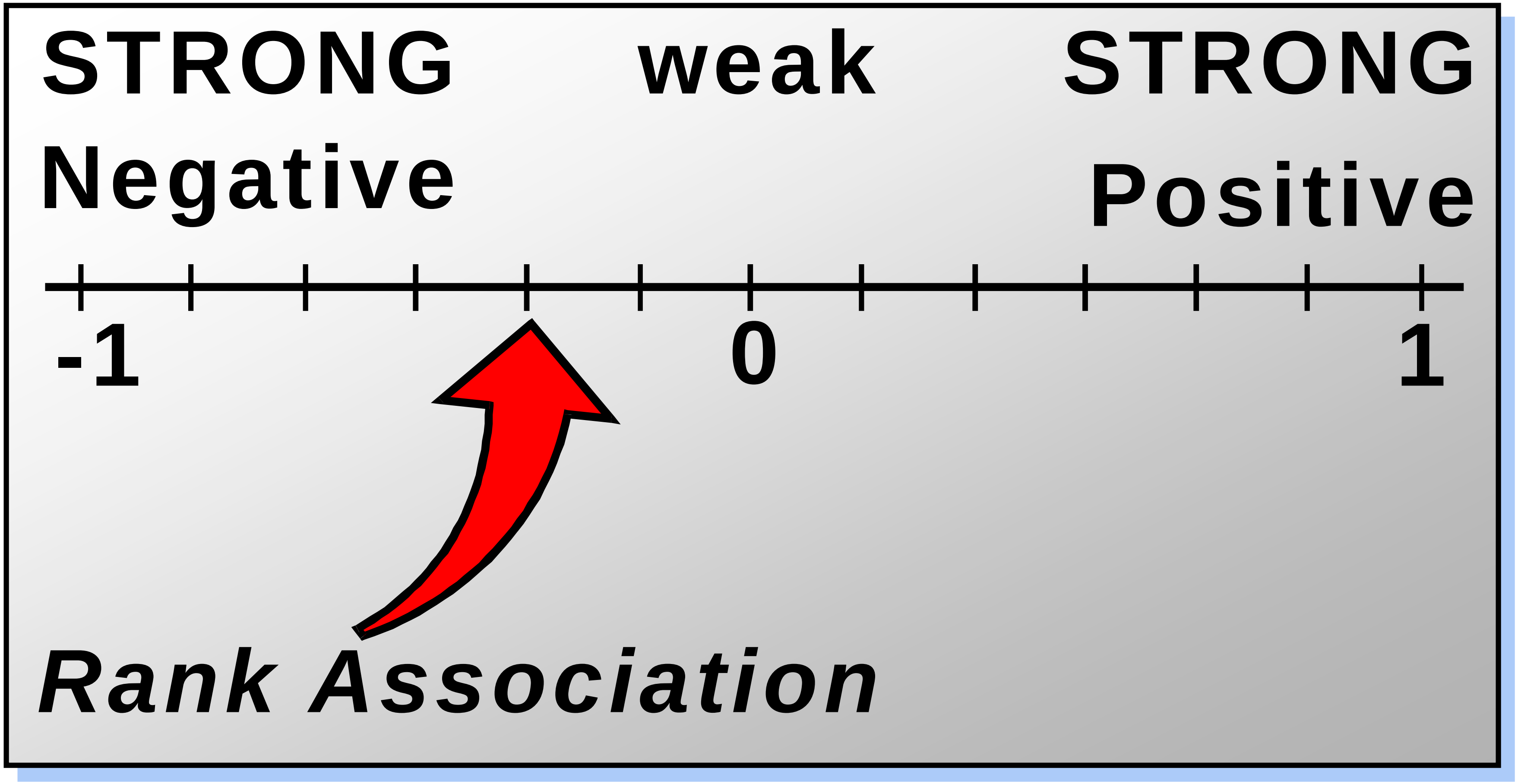


Test Ordinal Association

Mantel-Haenszel Chi-Square Test

- Determines whether an ordinal association exists
- DOES NOT measure the strength of the ordinal association
- Depends on and reflects the sample size

Spearman Correlation Statistic



Spearman versus Pearson

- The Spearman correlation uses ranks of the data.
- The Pearson correlation uses the observed values when the variable is numeric.

Detecting Ordinal Associations

```
proc freq data=bootcamp.ameshousing3;
    tables Fireplaces*Bonus / chisq measures cl;
    format Bonus bonusfmt.;
    title 'Ordinal Association between FIREPLACES and BONUS?';
run;
```

Ordinal Association between FIREPLACES and BONUS?
The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of Fireplaces by Bonus			
Fireplaces(Number of fireplaces)	Bonus(Sale Price > $175,000)		
	0	1	Total
0	177 59.00 90.77 69.41	18 6.00 9.23 40.00	195 65.00
1	68 22.67 73.12 26.67	25 8.33 26.88 55.56	93 31.00
2	10 3.33 83.33 3.92	2 0.67 16.67 4.44	12 4.00
Total	255 85.00	45 15.00	300 100.00

Detecting Ordinal Associations

Statistics for Table of Fireplaces by Bonus

Statistic	DF	Value	Prob
Chi-Square	2	15.4141	0.0004
Likelihood Ratio Chi-Square	2	14.4859	0.0007
Mantel-Haenszel Chi-Square	1	10.7458	0.0010
Phi Coefficient		0.2267	
Contingency Coefficient		0.2211	
Cramer's V		0.2267	

Detecting Ordinal Associations

Statistic	Value	ASE	95% Confidence Limits	
Gamma	0.4964	0.1111	0.2786	0.7143
Kendall's Tau-b	0.2072	0.0585	0.0926	0.3218
Stuart's Tau-c	0.1449	0.0433	0.0600	0.2298
Somers' D C|R	0.1510	0.0451	0.0626	0.2395
Somers' D R|C	0.2842	0.0786	0.1301	0.4383
Pearson Correlation	0.1896	0.0591	0.0737	0.3054
Spearman Correlation	0.2107	0.0594	0.0943	0.3272
Lambda Asymmetric C|R	0.0000	0.0000	0.0000	0.0000
Lambda Asymmetric R|C	0.0667	0.0603	0.0000	0.1849
Lambda Symmetric	0.0467	0.0424	0.0000	0.1298
Uncertainty Coefficient C|R	0.0571	0.0298	0.0000	0.1156
Uncertainty Coefficient R|C	0.0313	0.0167	0.0000	0.0640
Uncertainty Coefficient Symmetric	0.0404	0.0213	0.0000	0.0823

?

Poll

Quiz

Multiple Answer Poll

- A researcher wants to measure the strength of an association between two binary variables. Which statistic(s) can he use?
  a. Hansel and Gretel Correlation
  b. Mantel-Haenszel Chi-Square
  c. Pearson Chi-Square
  d. Odds Ratio
  e. Spearman Correlation

Multiple Answer Poll – Correct Answer

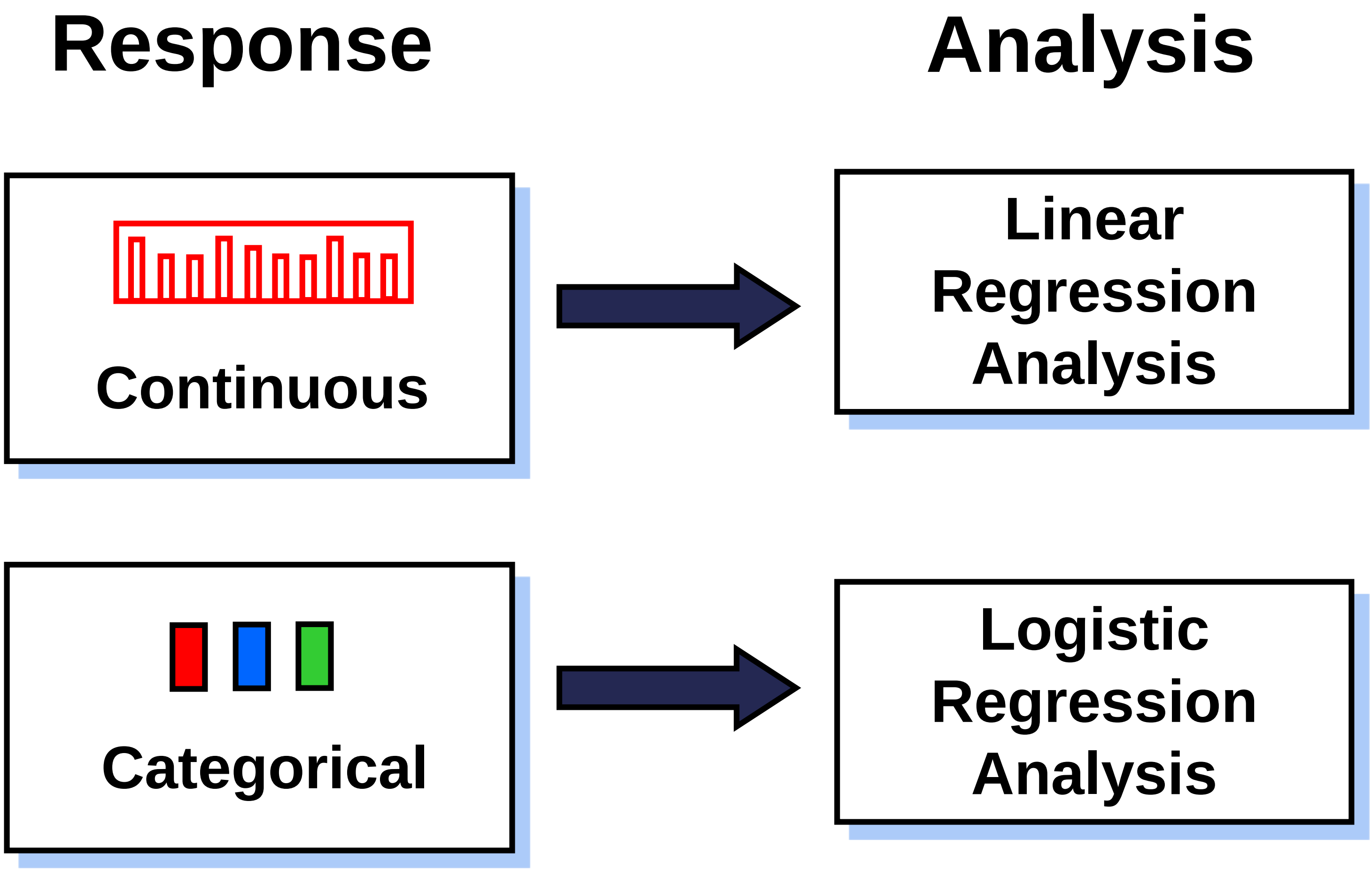
- A researcher wants to measure the strength of an association between two binary variables. Which statistic(s) can he use?
  a. Hansel and Gretel Correlation
  b. Mantel-Haenszel Chi-Square
  c. Pearson Chi-Square
  d. Odds Ratio
  e. Spearman Correlation

INTRODUCTION TO LOGISTIC REGRESSION

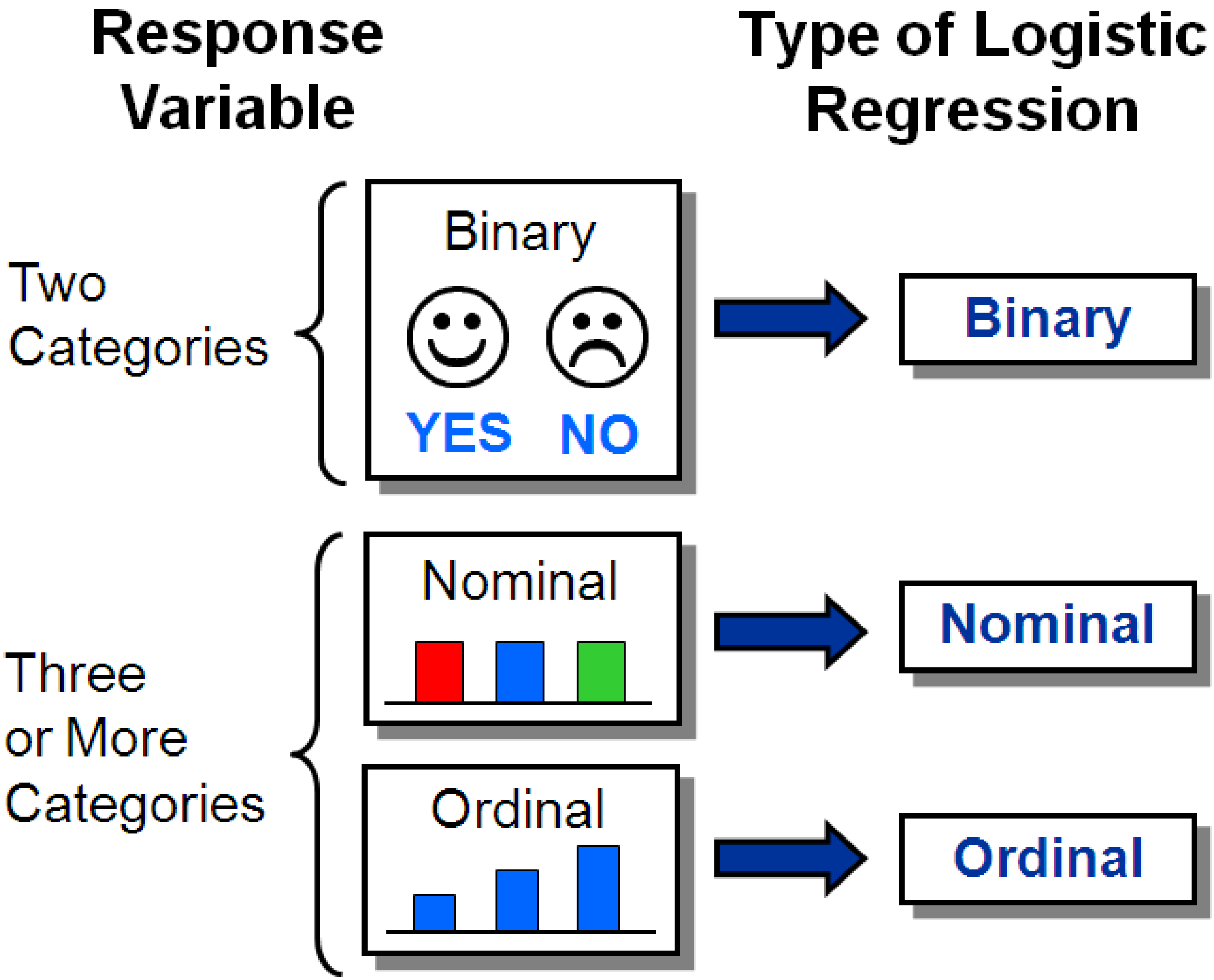
Objectives

- Define the concepts of logistic regression.
- Fit a binary logistic regression model using the LOGISTIC procedure.
- Describe the standard output from the LOGISTIC procedure with one continuous predictor variable.
- Read and interpret odds ratio tables and plots.

Overview


Response
Analysis
Continuous
Linear Regression Analysis
Categorical
Logistic Regression Analysis

Types of Logistic Regression



Why Not Least Squares Regression?

$$y_i = \beta_0 + \beta_1 x_{1,i} + \varepsilon_i$$

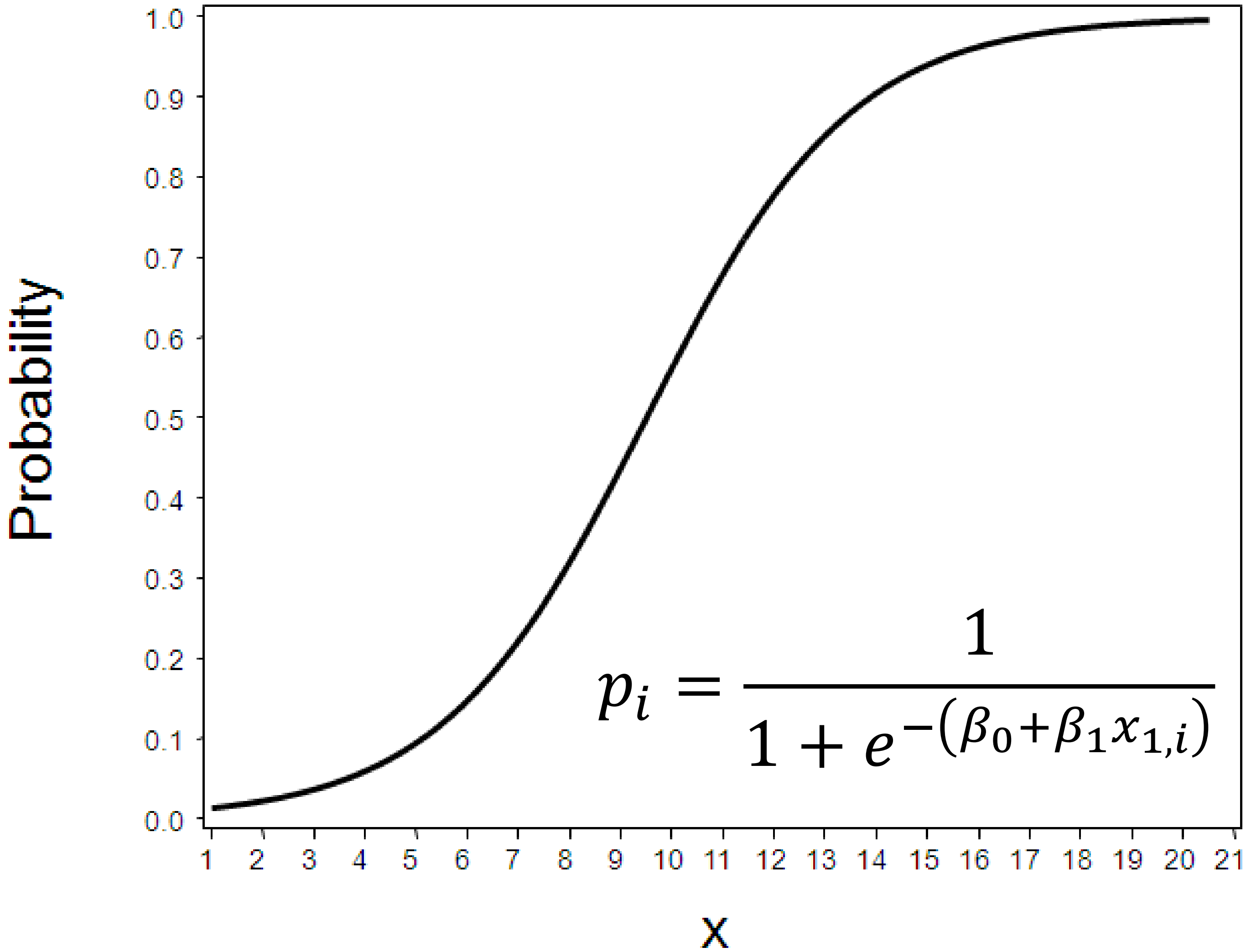
- If the response variable is categorical, then how do you code the response numerically?
- If the response is coded (1=Yes and 0=No) and your regression equation predicts 0.5 or 1.1 or -0.4, what does that mean practically?
- If there are only two (or a few) possible response levels, is it reasonable to assume constant variance and normality?

Linear Probability Model?

$$p_i = \beta_0 + \beta_1 x_{1,i}$$

- Probabilities are bounded, but linear functions can take on any value. (Once again, how do you interpret a predicted value of -0.4 or 1.1?)
- Given the bounded nature of probabilities, can you assume a linear relationship between X and *p* throughout the possible range of X?
- Can you assume a random error with constant variance?
- What is the observed probability for an observation?

Logistic Regression Model



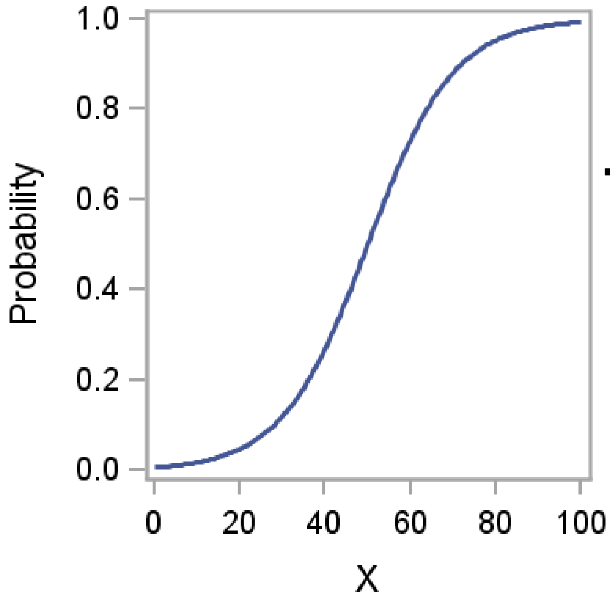
$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i})}}$$

The Logit Link Transformation

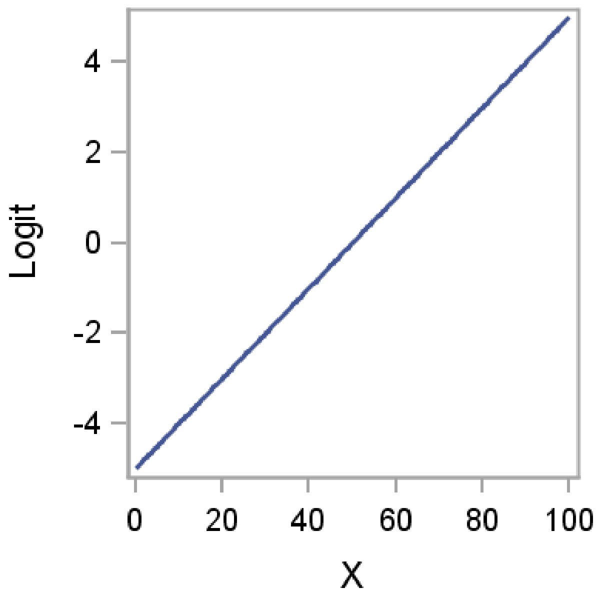
$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{1i}$$

- To create a linear model, a link function (logit) is applied to the probabilities.
- The relationship between the parameters and the logits are linear.
- Logits unbounded.

Assumption



Logit Transformation



Poll

Quiz

Multiple Choice Poll

- What are the upper and lower bounds for a logit?

a. Lower=0, Upper=1
b. Lower=0, No upper bound
c. No lower bound, No upper bound
d. No lower bound, Upper=1

Multiple Choice Poll – Correct Answer

- What are the upper and lower bounds for a logit?

a. Lower=0, Upper=1
b. Lower=0, No upper bound
c. No lower bound, No upper bound
d. No lower bound, Upper=1

Simple Logistic Regression Model

```
proc logistic data=bootcamp.ameshousing3 alpha=0.05
              plots(only)=(effect oddsratio);
    model Bonus(event='1')=Basement_Area / clodds=pl;
    title 'LOGISTIC MODEL (1):Bonus=Basement_Area';
run;
```

Simple Logistic Regression Model

LOGISTIC MODEL (1):Bonus=Basement_Area
The LOGISTIC Procedure

Model Information		
Data Set	BOOTCAMP.AMESHOUSING3	
Response Variable	Bonus	Sale Price > $175,000
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	300
Number of Observations Used	300

Simple Logistic Regression Model

Response Profile		
Ordered Value	**Bonus**	**Total Frequency**
1	0	255
2	1	45

Probability modeled is Bonus='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Simple Logistic Regression Model

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	255.625	161.838
SC	259.329	169.246
-2 Log L	253.625	157.838

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	95.7870	1	<.0001
Score	65.5624	1	<.0001
Wald	48.0617	1	<.0001

Simple Logistic Regression Model

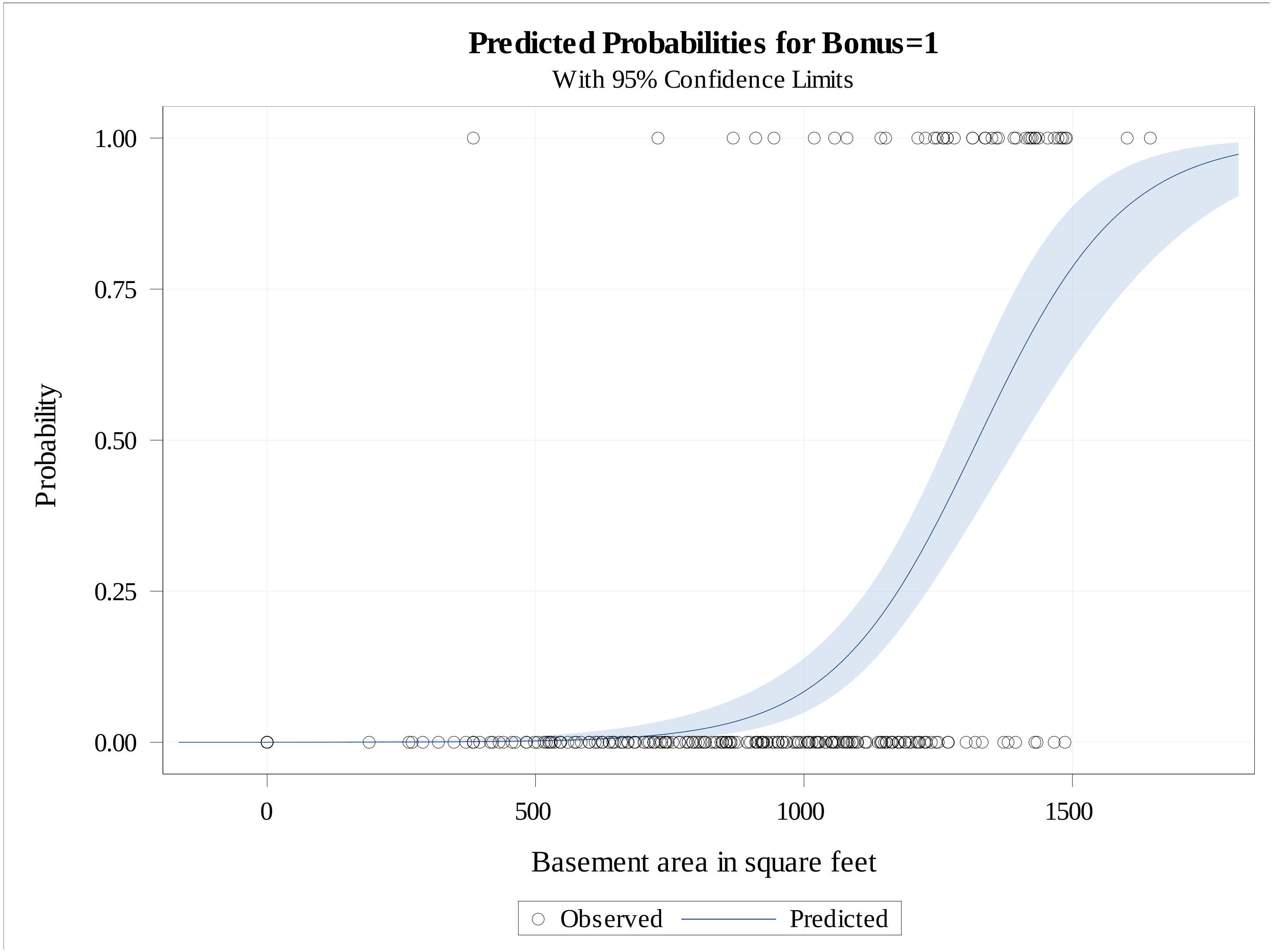
Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-9.7854	1.2896	57.5758	<.0001
Basement_ Area	1	0.00739	0.00107	48.0617	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.5	**Somers' D**	0.791
Percent Discordant	10.4	**Gamma**	0.792
Percent Tied	0.1	**Tau-a**	0.202
Pairs	11475	**c**	0.896

Simple Logistic Regression Model

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Basement_Area	1.0000	1.007	1.005	1.010

Simple Logistic Regression Model



Odds Ratio Calculation from the Current Logistic Regression Model

- Logistic regression model:

$$\operatorname{logit}\left(\hat{p}\right) = \log(odds) = \beta_0 + \beta_1 * \left(Basement\_Area\right)$$

- Odds ratio (1-year difference in Basement Area):

$$\text{odds}_{\text{larger}} = e^{\beta_0 + \beta_1 * (Basement\_Area + 1)}$$

$$\text{odds}_{\text{smaller}} = e^{\beta_0 + \beta_1 * (Basement\_Area)}$$

$$\text{Odds Ratio} = \frac{e^{\beta_0 + \beta_1 * (Basement\_Area + 1)}}{e^{\beta_0 + \beta_1 * (Basement\_Area)}} = e^{\beta_1}$$

$$= e^{(0.00739)} = 1.007$$

Odds Ratio for a Continuous Predictor

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Basement_Area	1.0000	1.007	1.005	1.010

Model Assessment: Comparing Pairs

- 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.

Comparing Pairs

To find concordant, discordant, and tied pairs, compare houses that had the outcome of interest against houses that did not.

Not Bonus Eligible

Bonus Eligible

Concordant Pair

Compare a 1200 square foot basement that was bonus eligible with an 800 square foot basement that was not.

Not Eligible, 800 sqft

P(Eligible)=.0204

Bonus Eligible, 1200 sqft

P(Eligible)=.2865

The actual sorting agrees with the model.

This is a **concordant** pair.

Discordant Pair

Compare a 1400 square foot basement that was bonus eligible with a 1600 square foot basement that was not.

Not Eligible, 1600 sq ft

P(Eligible)=.8855

Bonus Eligible, 1400 sq ft

P(Eligible)=.6379

The actual sorting disagrees with the model.

This is a **discordant** pair.

Tied Pair

Compare two 1350 square foot basements. One was bonus eligible and the other not.

Not Eligible, 1350 sqft

P(Eligible)=.5490

Bonus Eligible, 1350 sqft

P(Eligible)=.5490

The model cannot distinguish between the two.

This is a **tied** pair.

Concordant, Discordant, Tied Pairs

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.2	Somers' D	0.790
Percent Discordant	10.2	Gamma	0.795
Percent Tied	0.6	Tau-a	0.202
Pairs	11475	c	0.895

?

LOGISTIC REGRESSION WITH CATEGORICAL PREDICTORS

Objectives

- State how a logistic model with categorical predictors does and does not differ from one with continuous predictors.
- Describe what a CLASS statement does.
- Define the standard output from the LOGISTIC procedure with categorical predictor variables.

CLASS Statement

- The CLASS statement creates a set of “design variables” representing the information in the categorical variables.
- Character variables cannot be used, as is, in a model.
- The design variables are the ones actually used in model calculations.
- There are several “parameterizations” available in PROC LOGISTIC.

Effect (Default) Coding: Three Levels

			Design Variables	
CLASS	Value	Label	1	2
IncLevel	1	Low Income	1	0
	2	Medium Income	0	1
	3	High Income	-1	-1

Effect Coding: An Example

$$\text{logit}(p)=\beta_0+\beta_1*D_{\text{Low income}}+\beta_2*D_{\text{Medium income}}$$

β_0= the average value of the logit across all categories

β_1= the difference between the logit for Low income and the average logit

β_2= the difference between the logit for Medium income and the average logit

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-0.5363	0.1015	27.9143	<.0001
IncLevel	**1**	1	-0.2259	0.1481	2.3247	0.1273
IncLevel	**2**	1	-0.2200	0.1447	2.3111	0.1285

Reference Cell Coding: Three Levels

			Design Variables	
CLASS	Value	Label	1	2
IncLevel	1	Low Income	1	0
	2	Medium Income	0	1
	3	High Income	0	0

Reference Cell Coding: An Example

$$\text{logit}(p)=\beta_0+\beta_1{*}D_{\text{Low income}}+\beta_2{*}D_{\text{Medium income}}$$

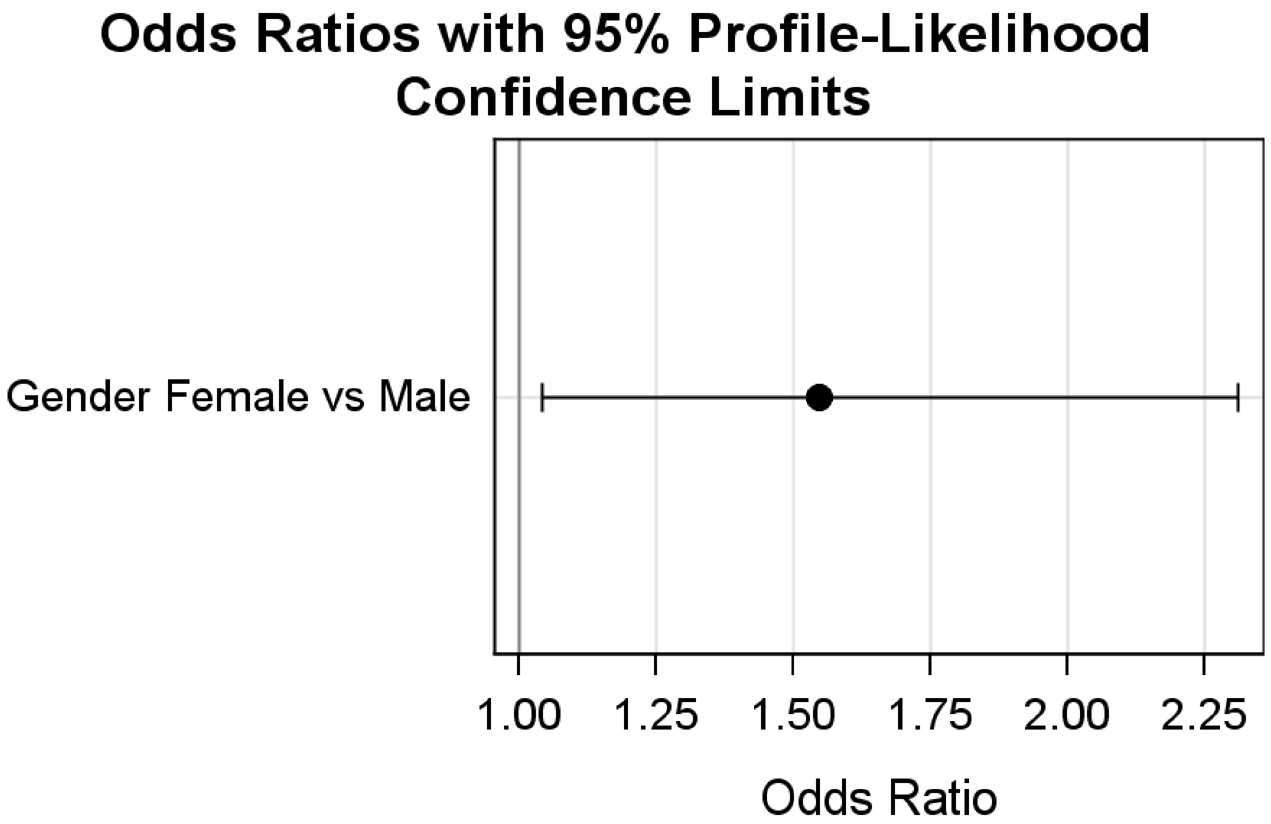
β_0= the value of the logit when income is High

β_1= the difference between the logits for Low and High income

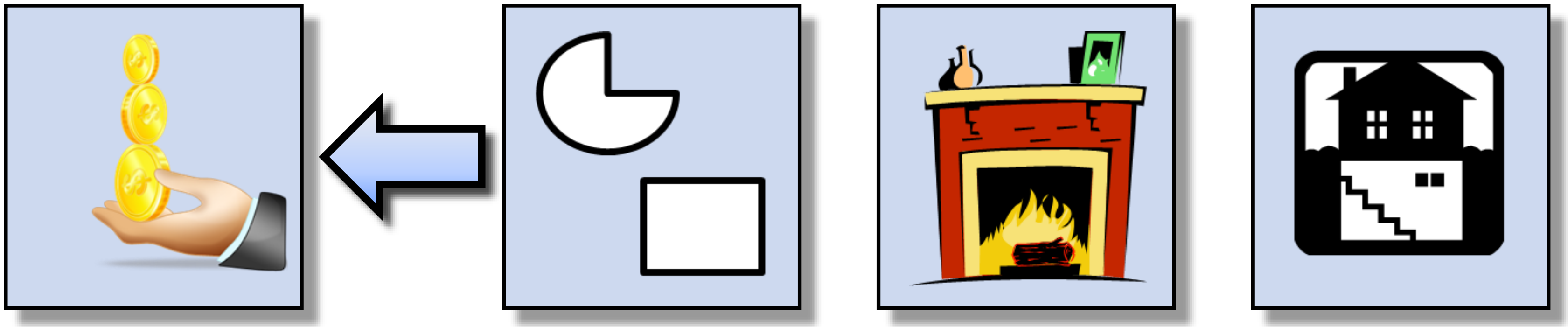
β_2= the difference between the logits for Medium and High income

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-0.0904	0.1608	0.3159	0.5741
IncLevel	**1**	1	-0.6717	0.2465	7.4242	0.0064
IncLevel	**2**	1	-0.6659	0.2404	7.6722	0.0056

Odds Ratio for Categorical Predictor



Multiple Logistic Regression



$$\text{logit}(p) = \beta_0 + \beta_1 X_{\text{irregular}} + \beta_2 X_{\text{fireplace=1}} + \beta_3 X_{\text{fireplace=2}} + \beta_4 X_{\text{Basement\_Area}}$$

Multiple Logistic Regression Model

```
proc logistic data=bootcamp.ameshousing3 plots(only)=(effect oddsratio);
    class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
    model Bonus(event='1')=Basement_Area Fireplaces Lot_Shape_2 / clodds=pl;
    units Basement_Area=100;
    title 'LOGISTIC MODEL (2):Bonus= Basement_Area Fireplaces Lot_Shape_2';
run;
```

Multiple Logistic Regression Model

Class Level Information			
Class	**Value**	**Design Variables**	
Fireplaces	**0**	0	0
	1	1	0
	2	0	1
Lot_Shape_2	**Irregular**	1	
	Regular	0	

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Multiple Logistic Regression Model

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	251.812	140.499
SC	255.513	159.001
-2 Log L	249.812	130.499

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	119.3133	4	<.0001
Score	91.7250	4	<.0001
Wald	49.8671	4	<.0001

Multiple Logistic Regression Model

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Basement_Area	1	38.1356	<.0001
Fireplaces	2	5.2060	0.0741
Lot_Shape_2	1	16.9421	<.0001

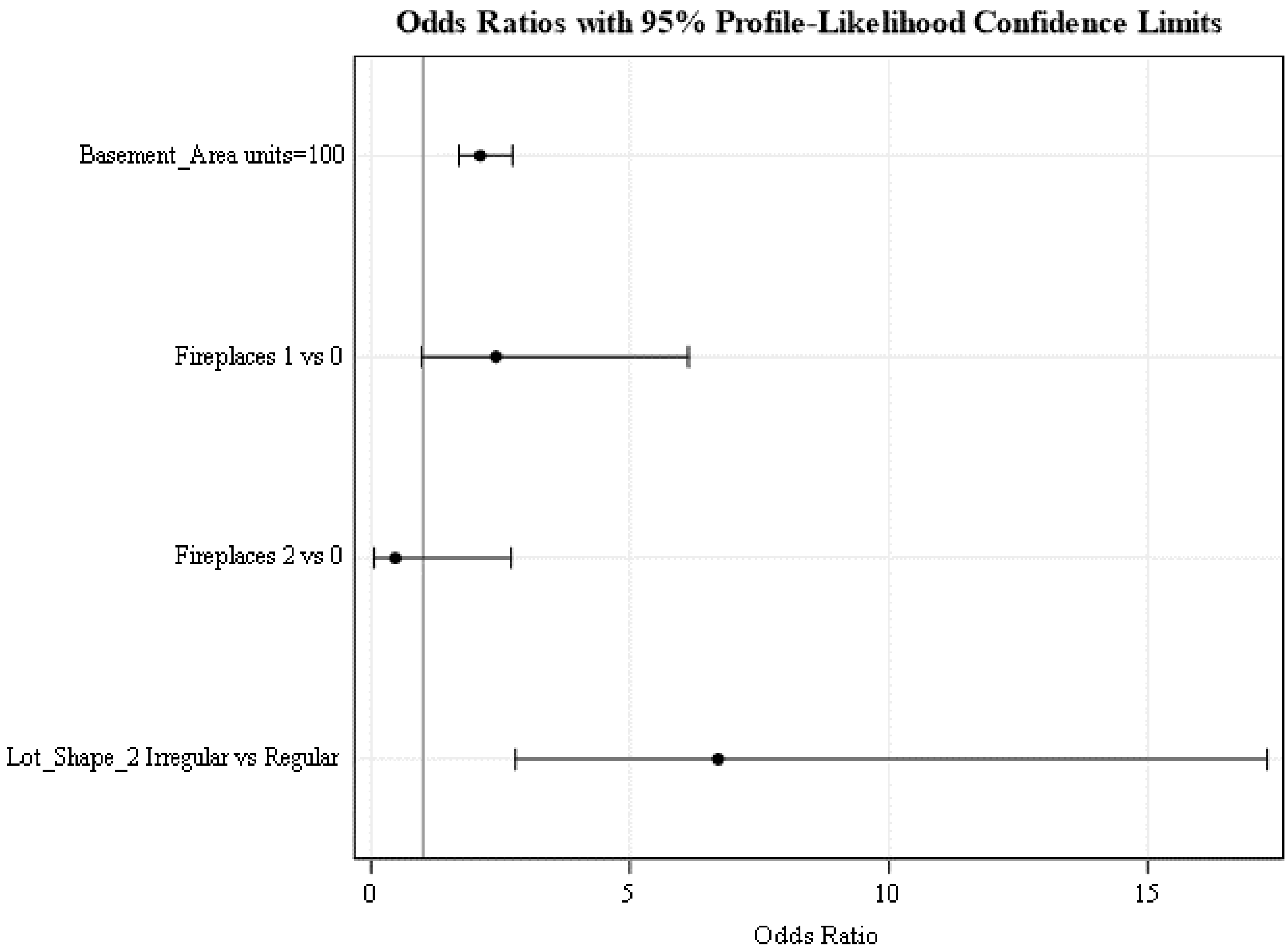
Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-11.0882	1.5384	51.9467	<.0001
Basement_Area		1	0.00744	0.00120	38.1356	<.0001
Fireplaces	**1**	1	0.8810	0.4658	3.5770	0.0586
Fireplaces	**2**	1	-0.7683	0.9654	0.6335	0.4261
Lot_Shape_2	**Irregular**	1	1.9025	0.4622	16.9421	<.0001

Multiple Logistic Regression Model

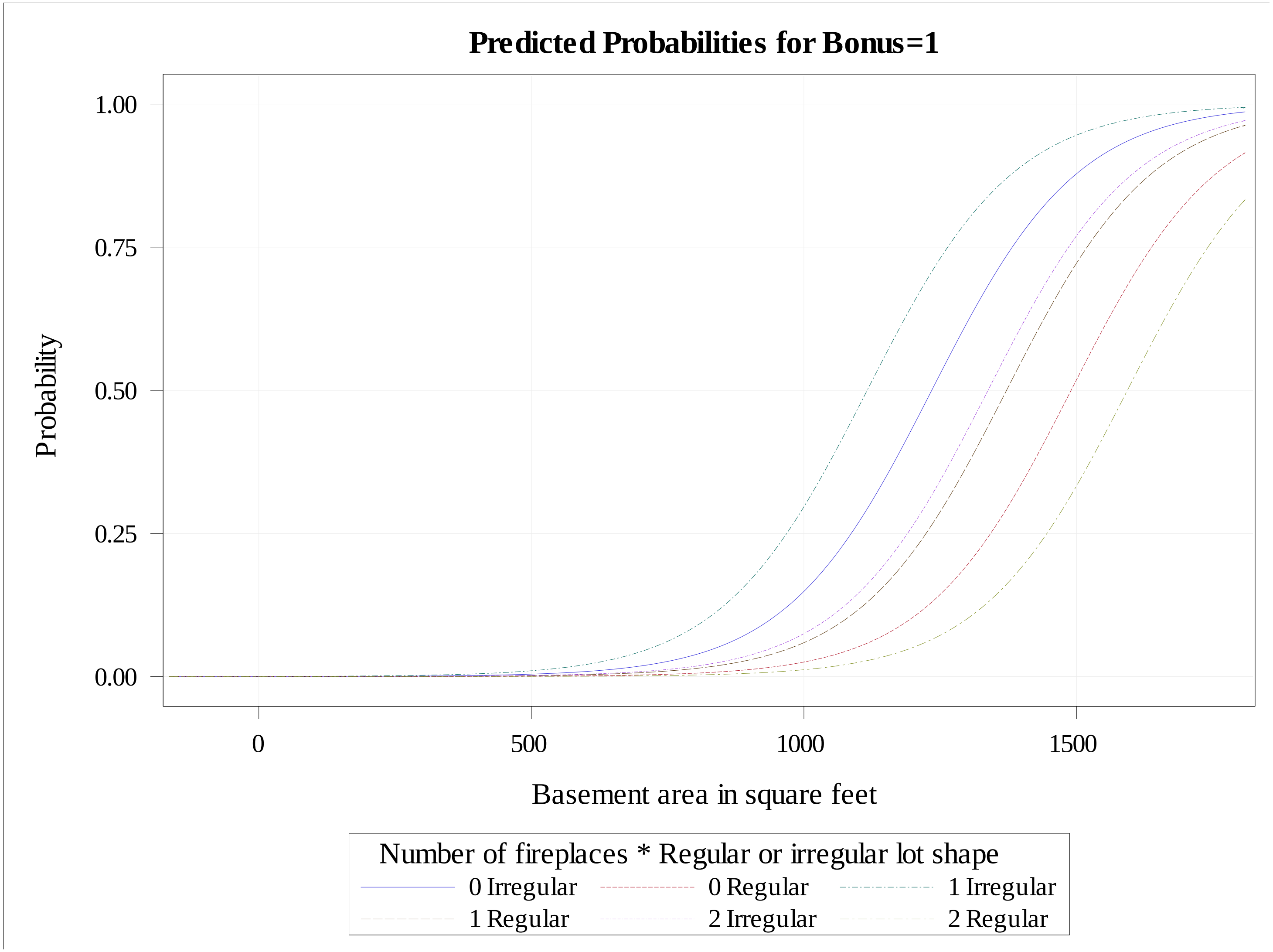
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	92.9	**Somers' D**	0.859
Percent Discordant	7.0	**Gamma**	0.860
Percent Tied	0.1	**Tau-a**	0.216
Pairs	11220	**c**	0.930

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	**Unit**	**Estimate**	**95% Confidence Limits**	
Basement_Area	100.0	2.105	1.696	2.727
Fireplaces 1 vs 0	1.0000	2.413	0.973	6.127
Fireplaces 2 vs 0	1.0000	0.464	0.054	2.703
Lot_Shape_2 Irregular vs Regular	1.0000	6.703	2.786	17.301

Multiple Logistic Regression Model



Multiple Logistic Regression Model



?

STEPWISE SELECTION WITH INTERACTIONS

Objectives

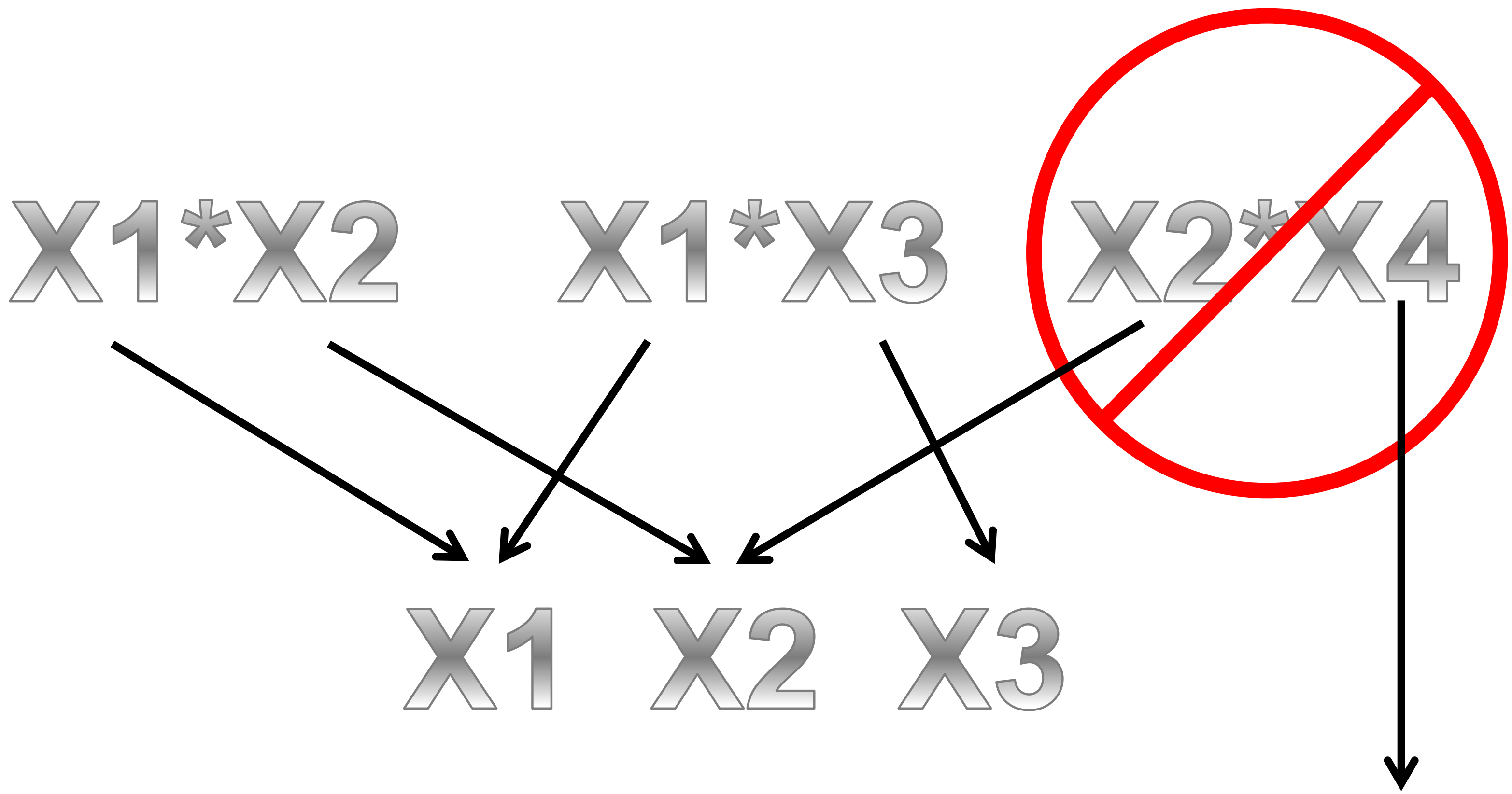
- Fit a multiple logistic regression model with main effects and interactions using the backward elimination method.
- Explain interactions using graphs.

Stepwise Methods – Default

	PROC REG/ PROC GLMSELECT			PROC LOGISTIC	
	SLENTRY	SLSTAY		SLENTRY	SLSTAY
FORWARD	0.50	-----		0.05	-----
BACKWARD	-----	0.10		-----	0.05
STEPWISE	0.15	0.15		0.05	0.05

Stepwise Hierarchy Rules

By default, at each step model hierarchy is retained. This means that higher level effects cannot be in a model when any of its lower level composite effects are not present.



Logistic Regression – Backward Selection

```
proc logistic data=bootcamp.ameshousing3 plots(only)=(effect oddsratio);
    class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
    model Bonus(event='1')=Basement_Area|Fireplaces|Lot_Shape_2 @2 /
          selection=backward clodds=pl slstay=0.10;
    units Basement_Area=100;
    title 'LOGISTIC MODEL (3): Backward Elimination '
           'Bonus=Basement_Area|Fireplaces|Lot_Shape_2';
run;
```

Logistic Regression – Predictions

```
data newhouses;
      length Lot_Shape_2 $9;
      input Fireplaces Lot_Shape_2 $ Basement_Area;
      datalines;
      0   Regular     1060
      2   Regular      775
      2   Irregular   1100
      1   Irregular    975
      1   Regular      800
      ;
run;

proc logistic data=bootcamp.ameshousing3;
    class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
    model Bonus(event='1')=Basement_Area|Lot_Shape_2 Fireplaces;
    units Basement_Area=100;
    score data=newhouses out=scored_houses;
run;
```

Logistic Regression – Predictions

VIEWTABLE: Work.Scored_houses (Posterior Probabilities for DATA=WORK.NEWHOUSES.)

	Lot_Shape_2	Fireplaces	Basement_Area	Into: Bonus	Predicted Probability: Bonus=0	Predicted Probability: Bonus=1
1	Regular	0	1060	0	0.98	0.022
2	Regular	2	775	0	1	4E-4
3	Irregular	2	1100	0	0.86	0.142
4	Irregular	1	975	0	0.69	0.306
5	Regular	1	800	0	1	0.003

?