



Categorical Data Analysis

Institute for Advanced Analytics
MSA Class of 2025

Qualitative Data Types

- **Categorical Variables:**

- Data whose measurement scale is inherently categorical.
- **Nominal** – categories with no logical ordering
- **Ordinal** – categories with a logical order / only two ways to order the categories (binary IS ordinal)

Ames Housing Data

```
train <- train %>%  
  mutate(Bonus = ifelse(Sale_Price > 175000, 1, 0))
```

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

	Bonus Eligible	
	Yes	No
Central Air	41%	59%
No Central Air	41%	59%

Association

	Bonus Eligible	
	Yes	No
Central Air	44%	56%
No Central Air	3%	97%

Exploring the Data

```
table(train$Central_Air)
```

```
##
```

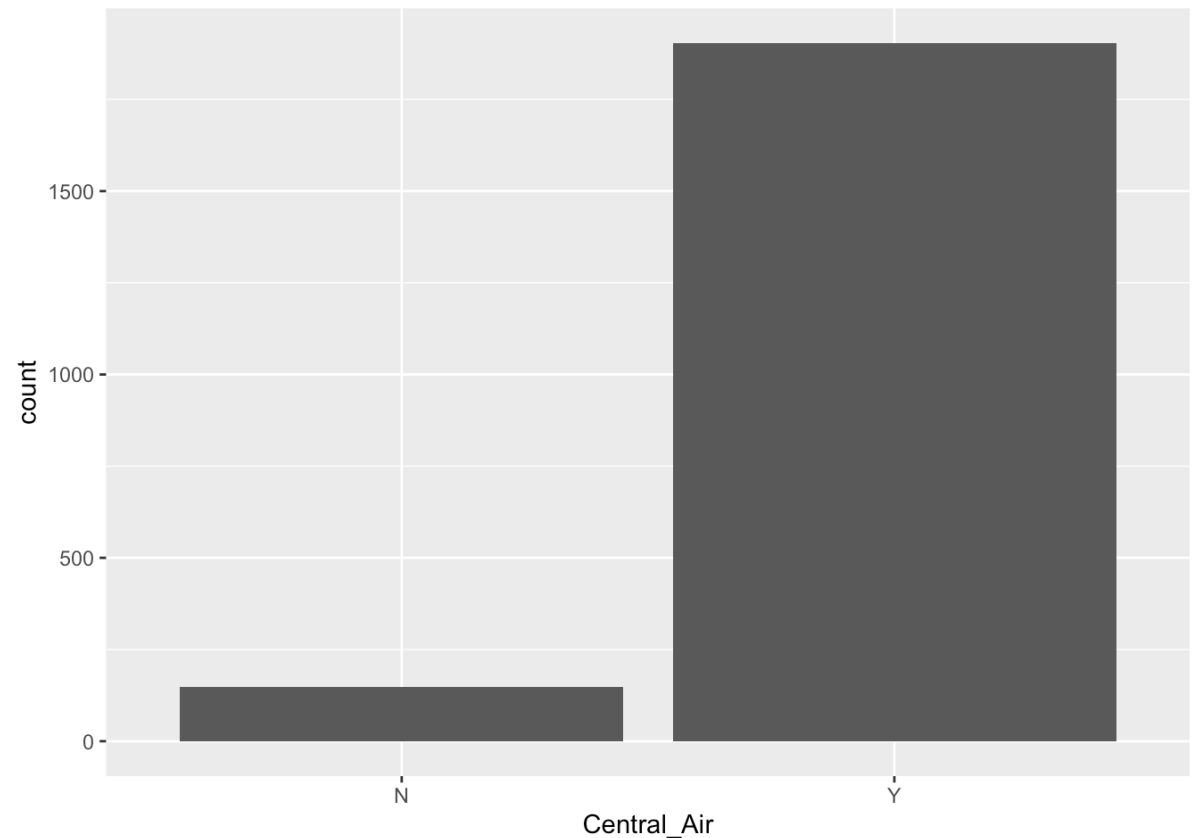
```
##      N      Y
```

```
##  147 1904
```

```
ggplot(data = train) +
```

```
  geom_bar(mapping =
```

```
    aes(x = Central_Air))
```



Exploring the Data

```
table(train$Bonus)
```

```
##
```

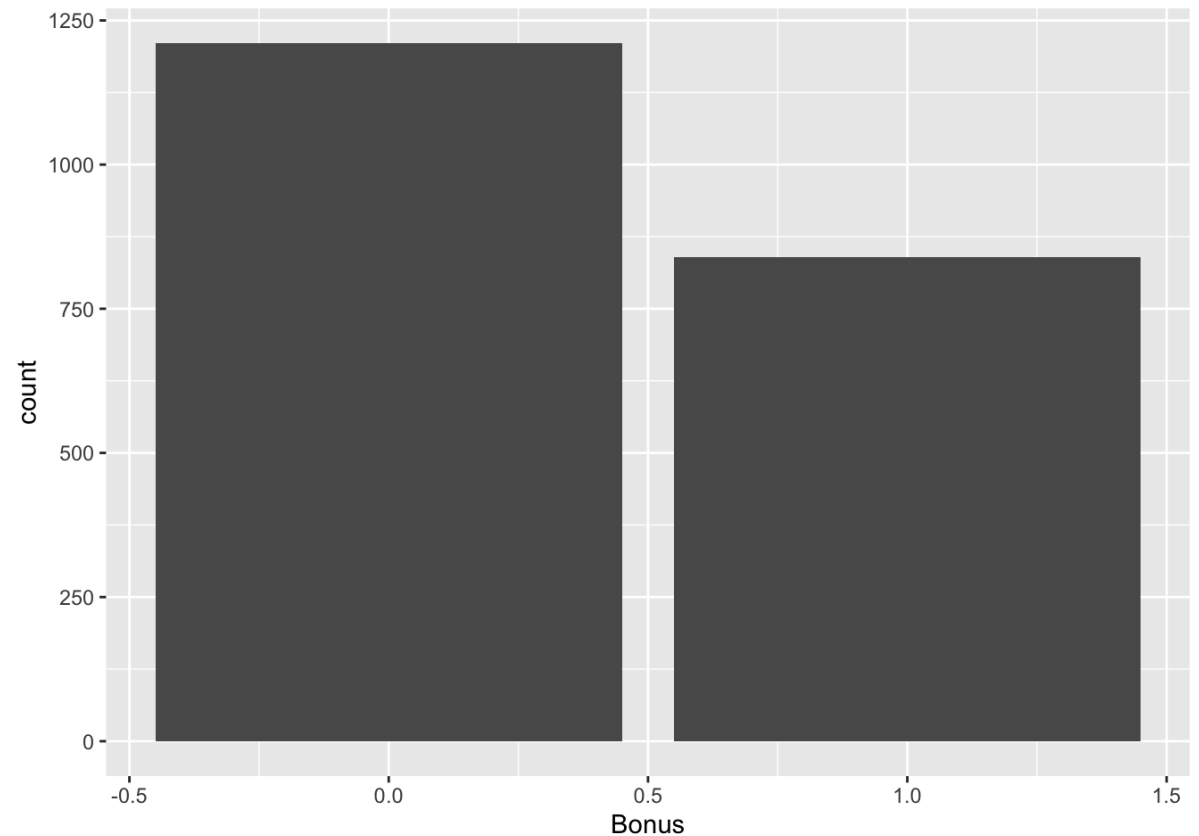
```
##      0      1
```

```
## 1211 840
```

```
ggplot(data = train) +
```

```
  geom_bar(mapping =
```

```
    aes(x = Bonus))
```



Cross-Tabulation Table

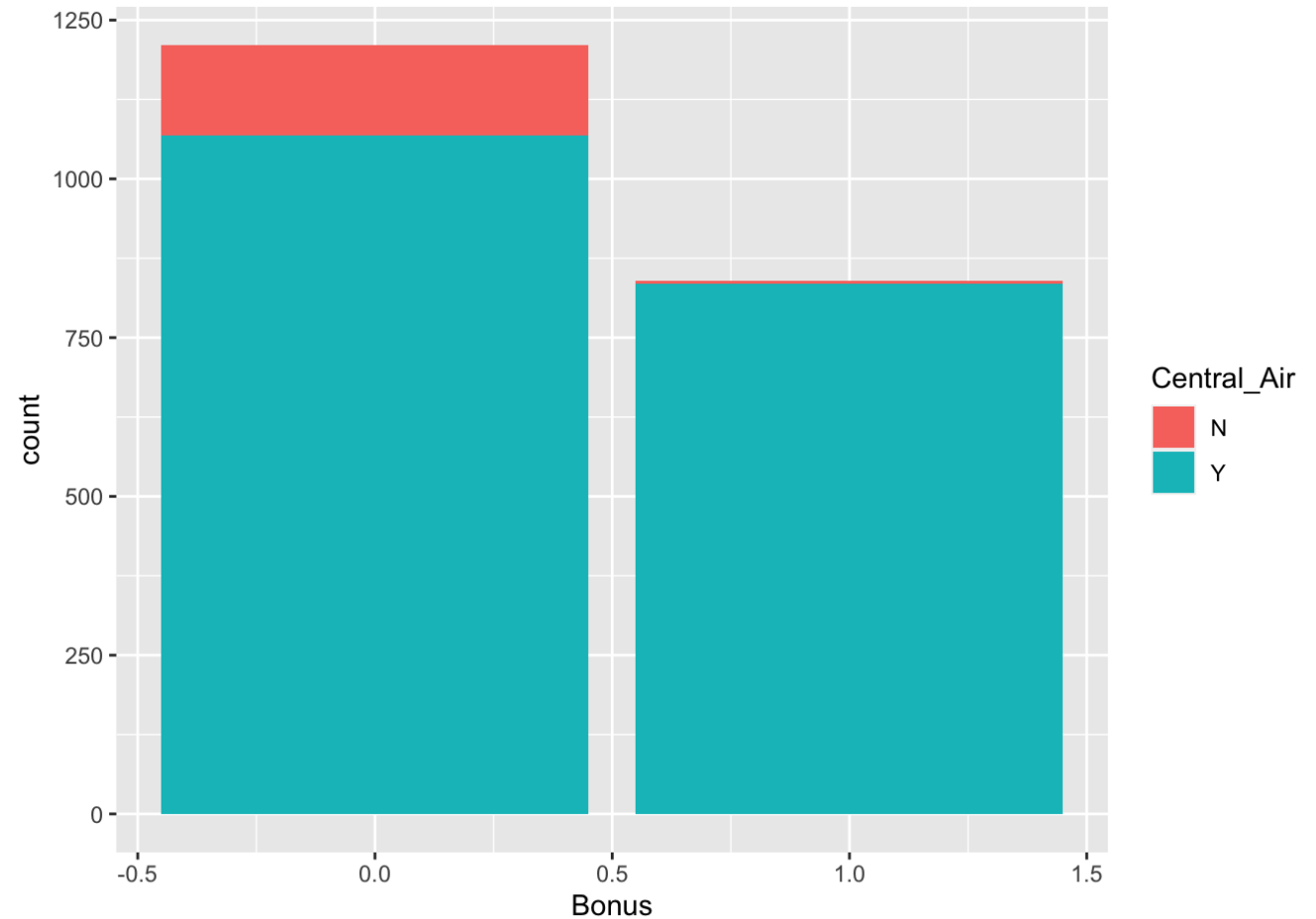
- Explore two variables with **cross-tabulation** tables.
- Shows the number of observations for each combination of the row and column variables.

Exploring the Data

```
table(train$Central_Air, train$Bonus)
```

```
##  
##      0      1  
## N   142     5  
## Y  1069   840
```

```
ggplot(data = train) +  
  geom_bar(mapping =  
    aes(x = Bonus,  
        fill = Central_Air))
```



Exploring the Data

```
library(gmodels)
```

```
CrossTable(train$Central_Air, train$Bonus)
```

Exploring the Data

##	Cell Contents	##	train\$Bonus			
##	-----	##	train\$Central_Air	0	1	Row Total
##	N	##	-----	-----	-----	-----
##	Chi-square contribution	##	N	142	5	147
##	N / Row Total	##		35.112	50.620	
##	N / Col Total	##		0.966	0.034	0.072
##	N / Table Total	##		0.117	0.006	
##	-----	##		0.069	0.002	
		##	-----	-----	-----	-----
		##	Y	1069	835	1904
		##		2.711	3.908	
		##		0.561	0.439	0.928
		##		0.883	0.994	
		##		0.521	0.407	
		##	-----	-----	-----	-----
		##	Column Total	1211	840	2051
		##		0.590	0.410	
		##	-----	-----	-----	-----



Tests of Association

Association

	Bonus Eligible	
	Yes	No
Central Air	44%	56%
No Central Air	3%	97%

How **much of a change** is required
to believe there actually is a difference?

Tests of Association - Hypotheses

- **Null Hypothesis**

- There is no association.
- The distribution of one variable does not change across levels of another variable.

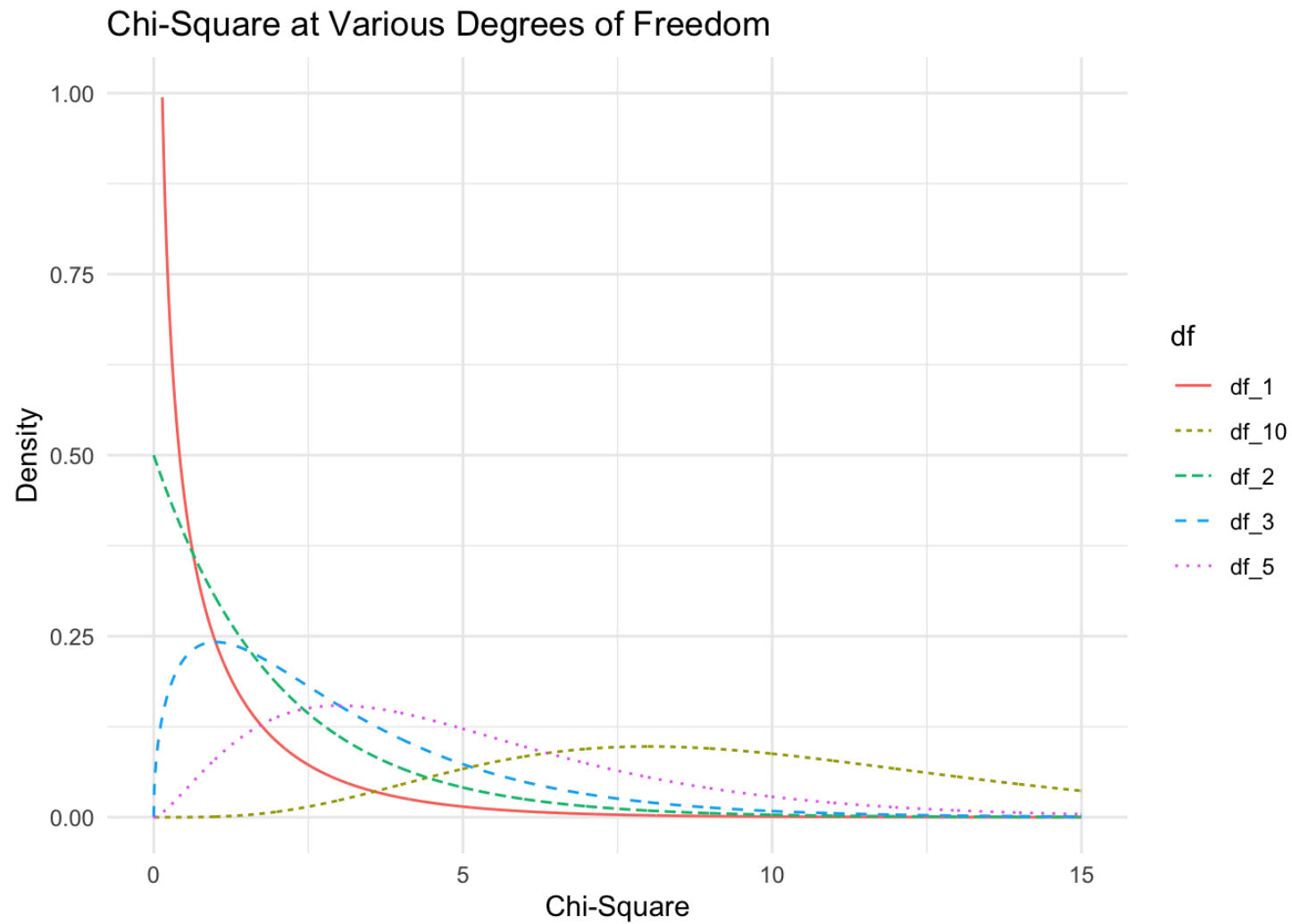
- **Alternative Hypothesis**

- There *is* an association.
- The distribution of one variable changes across levels of another variable.

χ^2 -Distribution

- The Chi-Square test comes from the χ^2 -**distribution**.
- Characteristics of the χ^2 -distribution:
 1. Bounded Below By Zero
 2. Right Skewed
 3. One set of Degrees of Freedom

χ^2 -Distribution



Pearson Chi-Square Test

- The Pearson χ^2 test works for comparing any two categorical variables.

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

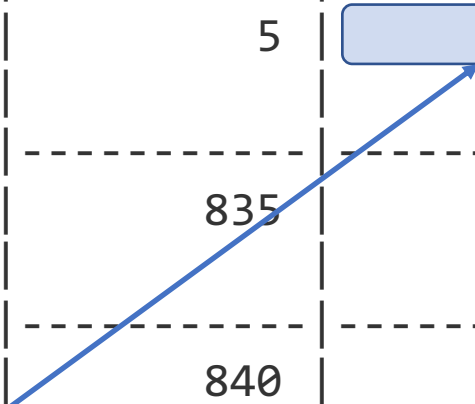
$$\text{D.F.} = (\# \text{ Rows} - 1)(\# \text{ Columns} - 1)$$

Expected Cell Counts

##		train\$Bonus		
## train\$Central_Air		0	1	Row Total
## -----		-----	-----	-----
## N		142	5	147
##				
## -----		-----	-----	-----
## Y		1069	835	1904
##				
## -----		-----	-----	-----
## Column Total		1211	840	2051
##		0.590	0.410	
## -----		-----	-----	-----

Expected Cell Counts

##		train\$Bonus		
## train\$Central_Air		0	1	Row Total
## -----		-----	-----	-----
##	N	142	5	147
##				
##	-----	-----	-----	-----
##	Y	1069	835	1904
##				
##	-----	-----	-----	-----
##	Column Total	1211	840	2051
##		0.590	0.410	
##	-----	-----	-----	-----



Expected Cell Counts

##		train\$Bonus		
## train\$Central_Air		0	1	Row Total
## -----		-----	-----	-----
##	N	142	5	147
##		86.73	60.27	
## -----		-----	-----	-----
##	Y	1069	835	1904
##				
## -----		-----	-----	-----
##	Column Total	1211	840	2051
##		0.590	0.410	
## -----		-----	-----	-----

Expected Cell Counts

##		train\$Bonus		
## train\$Central_Air		0	1	Row Total
## -----		-----	-----	-----
## N		142	5	147
##		86.73	60.27	
## -----		-----	-----	-----
## Y		1069	835	1904
##		1123.36	780.64	
## -----		-----	-----	-----
## Column Total		1211	840	2051
##		0.590	0.410	
## -----		-----	-----	-----

Pearson Chi-Square Test

```
chisq.test(table(train$Central_Air, train$Bonus))
```

```
##
```

```
## Pearson's Chi-squared test with Yates' continuity correction
```

```
##
```

```
## data:  table(train$Central_Air, train$Bonus)
```

```
## X-squared = 90.686, df = 1, p-value < 2.2e-16
```

Likelihood Ratio Chi-Square Test

- The Likelihood Ratio χ^2 test works for comparing any two categorical variables.

$$\chi_{LR}^2 = 2 \times \sum_{i=1}^R \sum_{j=1}^C Obs_{i,j} \times \log \left(\frac{Obs_{i,j}}{Exp_{i,j}} \right)$$

$$D.F. = (\# \text{ Rows} - 1)(\# \text{ Columns} - 1)$$

Assumptions

- Both of the above tests have a sample size requirement.
- The sample size requirement is 80% or more of the cells in the cross-tabulation table need **expected** count larger than 5.

##		train\$Bonus		
##	train\$Central_Air	0	1	Row Total
##	-----	-----	-----	-----
##	N	142	5	147
##		86.73	60.27	
##	-----	-----	-----	-----
##	Y	1069	835	1904
##		1123.36	780.64	
##	-----	-----	-----	-----
##	Column Total	1211	840	2051
##		0.590	0.410	
##	-----	-----	-----	-----

Fisher's Exact Test

- When we **don't** meet the assumption, we can use the **Fisher's exact test** that calculates all possible permutations of data.

```
fisher.test(table(train$Central_Air, train$Bonus))
```

```
##  
## Fisher's Exact Test for Count Data  
##  
## data:  table(train$Central_Air, train$Bonus)  
## p-value < 2.2e-16
```

Ordinal Compared to Nominal Tests

- Both the Pearson and Likelihood Ratio Chi-Square tests can handle any type of categorical variable – either ordinal, nominal, or both.
- However, ordinal variables provide us extra information since the order of the categories actually matters compared to nominal.
- We can test for even more with ordinal variables against other ordinal variables – whether two ordinal variables have a **linear relationship** as compared to just a general one.

Mantel-Haenszel Chi-Square Test

- The Mantel-Haenszel χ^2 test works for comparing any two **ordinal** variables.

$$\chi_{MH}^2 = (n - 1)r^2$$

$$\text{D.F.} = 1$$

Mantel-Haenszel Chi-Square Test

```
library(vcdExtra)
```

```
CMHtest(table(train$Central_Air, train$Bonus))$table[1,]
```

```
##           Chisq           Df          Prob  
## 9.230619e+01 1.000000e+00 7.425180e-22
```





Measures of Association

Chi-Square Tests

- Determines whether an association exists
- **MAY NOT** measure the strength of the association
 - Can compare when sample size similar
 - Can **NOT** compare when sample size **different**

Measures of Association

- Measures the **strength of the association**
- There are many different measures of association.
- Two common measures of association are the following:
 1. Odds Ratios (Only for 2x2 tables – binary vs. binary)
 2. Cramer's V (Any size table)
 3. Spearman's Correlation (ordinal vs. ordinal)

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.
- The **odds** of an event occurring is NOT the same as the probability that an event occurs.

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.
- The **odds** of an event occurring is NOT the same as the probability that an event occurs.

$$Odds = \frac{p}{1 - p}$$

Probability versus Odds of an Outcome

##		train\$Bonus		
## train\$Central_Air		0	1	Row Total
## -----		-----	-----	-----
## N		142	5	147
##		0.966	0.034	
## -----		-----	-----	-----
## Y		1069	835	1904
##		0.561	0.439	
## -----		-----	-----	-----
## Column Total		1211	840	2051
##		0.590	0.410	
## -----		-----	-----	-----

Probability of **NOT** bonus eligible without central air = 0.966

Probability versus Odds of an Outcome

##		train\$Bonus		
## train\$Central_Air		0	1	Row Total
## -----		-----	-----	-----
## N		142	5	147
##		0.966	0.034	
## -----		-----	-----	-----
## Y		1069	835	1904
##		0.561	0.439	
## -----		-----	-----	-----
## Column Total		1211	840	2051
##		0.590	0.410	
## -----		-----	-----	-----

Odds of **NOT**
bonus eligible
without central
air

Probability of **NOT** bonus eligible without central air = 0.966

Probability of bonus eligible without central air = 0.034

$$= \frac{0.966}{0.034} = 28.41$$

Odds Ratio

##		train\$Bonus		
## train\$Central_Air		0	1	Row Total
## -----		-----	-----	-----
## N		142	5	147
##		0.966	0.034	
## -----		-----	-----	-----
## Y		1069	835	1904
##		0.561	0.439	
## -----		-----	-----	-----
## Column Total		1211	840	2051
##		0.590	0.410	
## -----		-----	-----	-----

Odds of **NOT** bonus eligible without central air = 28.41

Odds of **NOT** bonus eligible with central air = 1.28

Odds ratio

$$= \frac{28.41}{1.28} = 22.2$$

Odds Ratio

$$\begin{array}{l} \text{Odds of **NOT** bonus eligible without central air} = 28.41 \\ \text{Odds of **NOT** bonus eligible with central air} = 1.28 \end{array} \quad \begin{array}{l} \text{Odds ratio} \\ = \frac{28.41}{1.28} = 22.2 \end{array}$$

- Homes without central air have **22.2 times the odds** (22.2 times as likely) to not be bonus eligible as compared to homes with central air.

Odds Ratio

Odds of **NOT** bonus eligible without central air = 28.41

Odds of **NOT** bonus eligible with central air = 1.28

Odds ratio

$$= \frac{28.41}{1.28} = 22.2$$

- Homes without central air have **22.2 times the odds** (22.2 times as likely) to not be bonus eligible as compared to homes with central air.
- Reverse is also true!
- Homes with central air are 22.2 times as likely to be bonus eligible as compared to homes without central air.

Odds Ratio

```
library(DescTools)
OddsRatio(table(train$Central_Air, train$Bonus))

## [1] 22.18335
```

Cramer's V

- Odds ratios provide value for binary vs. binary relationships, but when you have more than two categories in one or both variables use **Cramer's V**.

$$V = \sqrt{\frac{\left(\frac{\chi_P^2}{n}\right)}{\min(\#Rows - 1, \#Columns - 1)}}$$

- Bounded between 0 and 1 (-1 and 1 for 2x2 scenario) where closer to 0 the weaker the relationship.

Cramer's V

```
assocstats(table(train$Central_Air, train$Bonus))
```

```
##              X^2 df P(> X^2)
## Likelihood Ratio 121.499  1      0
## Pearson          92.351  1      0
##
## Phi-Coefficient   : 0.212
## Contingency Coeff.: 0.208
## Cramer's V        : 0.212
```

Spearman's Correlation

- Spearman's correlation measures the strength of association between two ordinal variables.
- Calculated with the Pearson's correlation on the ranks of the observations instead of the values of the observations.

Spearman's Correlation

```
cor.test(x = as.numeric(ordered(train$Central_Air)),  
         y = as.numeric(ordered(train$Bonus)),  
         method = "spearman")
```

```
## Spearman's rank correlation rho  
##  
## data:  x and y  
## S = 1132826666, p-value < 2.2e-16  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
##          rho  
## 0.2121966
```



Introduction to Logistic Regression

Modeling Categorical Data

**Continuous Target
Variable**



**Linear Regression
ANOVA
Regularized Regression**

**Categorical Target
Variable**



Logistic Regression

Modeling Categorical Data

**Categorical Target
Variable**

Logistic Regression

Binary

Binary

Ordinal

Ordinal

Nominal

Nominal

Introduction to Logistic Regression

LINEAR PROBABILITY MODEL

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

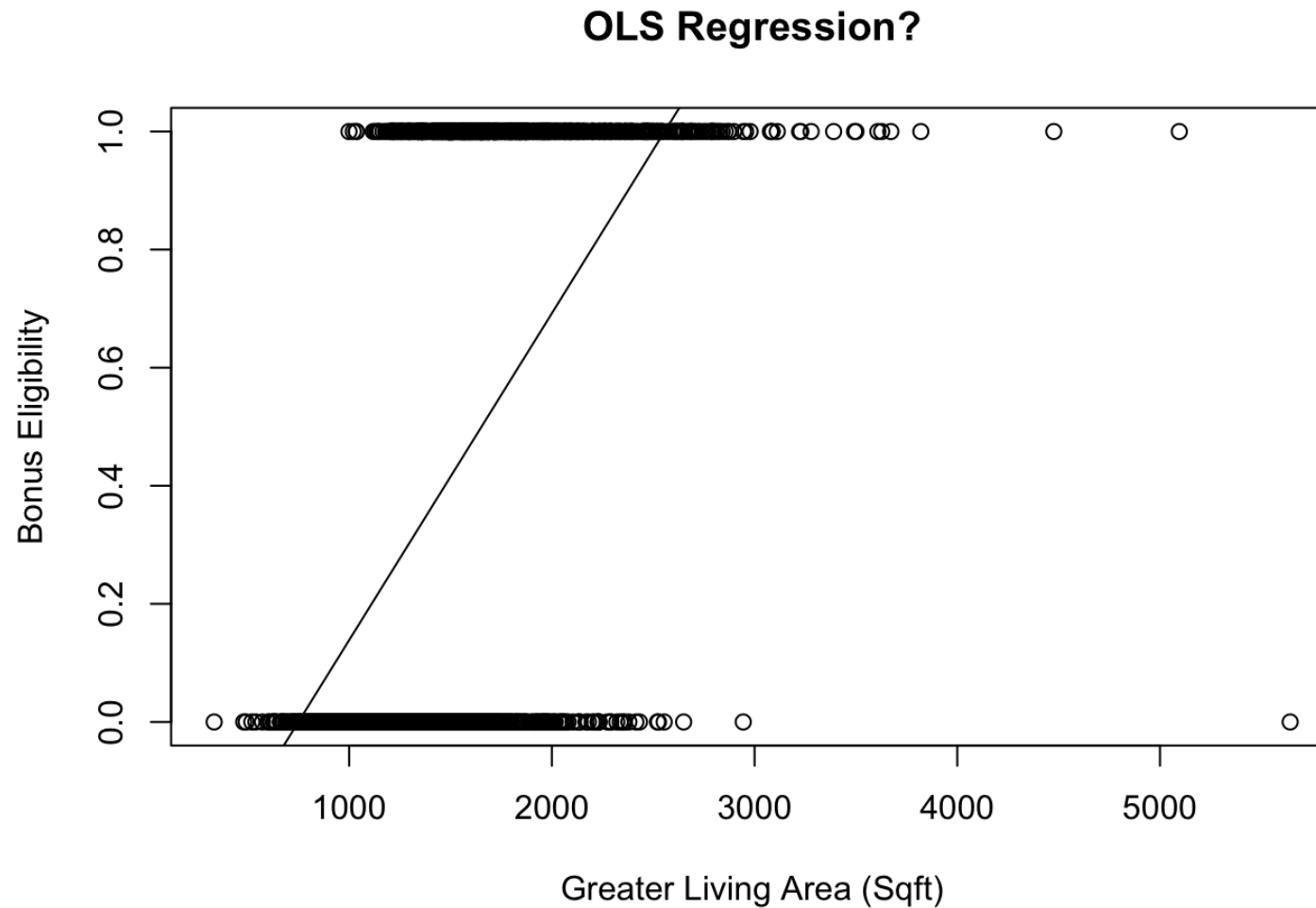
$$p_i = \beta_0 + \beta_1 x_{1,i}$$

- 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?

Linear Probability Model

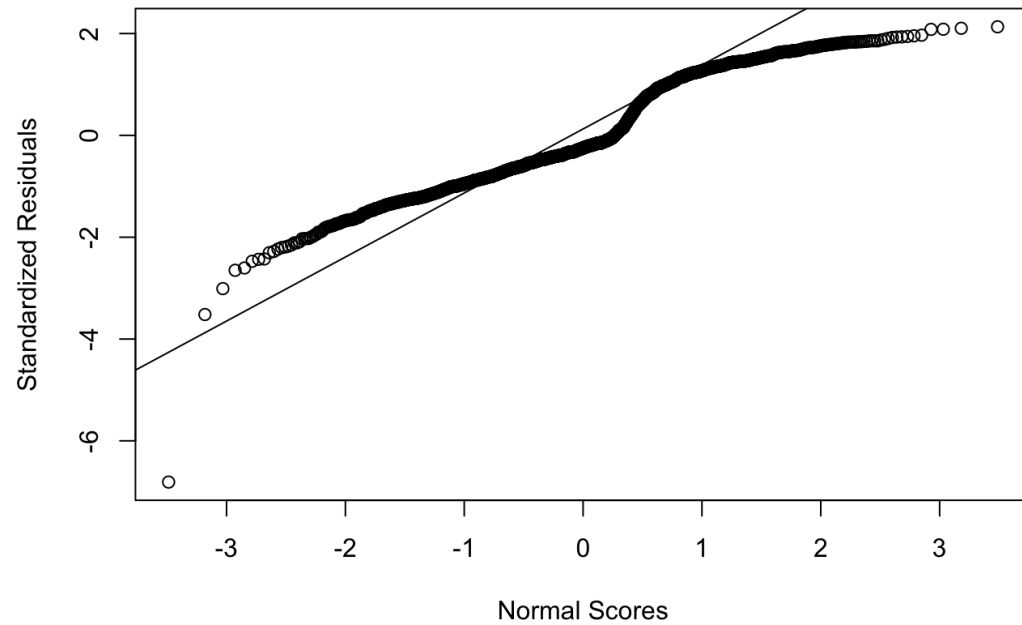
```
lp.model <- lm(Bonus ~ Gr_Liv_Area, data = train)
with(train, plot(x = Gr_Liv_Area, y = Bonus,
                 main = 'OLS Regression?',
                 xlab = 'Greater Living Area (Sqft)',
                 ylab = 'Bonus Eligibility'))
abline(lp.model)
```

Linear Probability Model

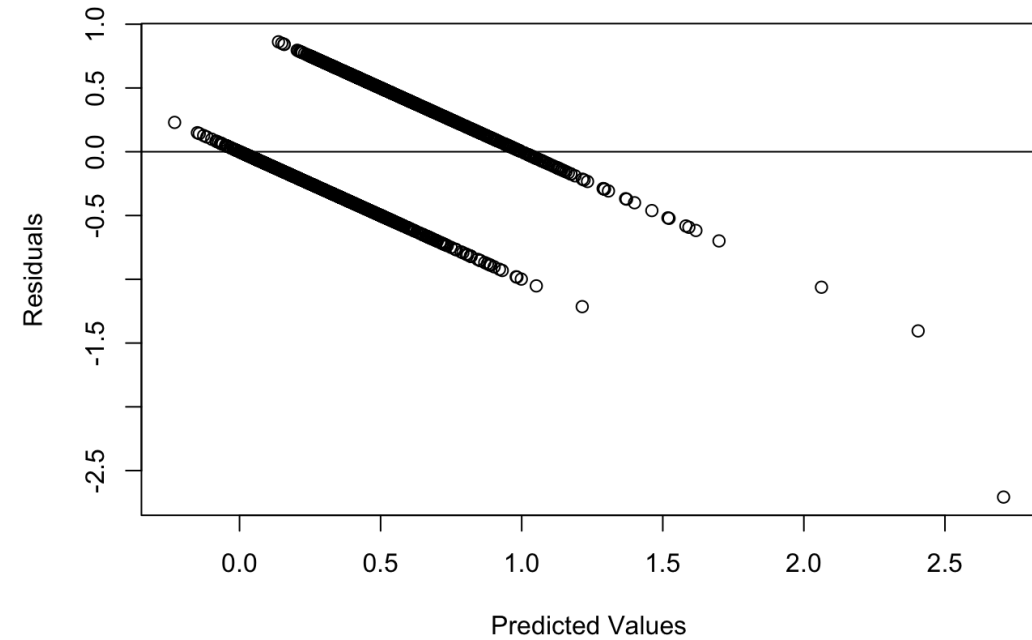


Linear Probability Model

QQ-Plot of Residuals



Residuals of Linear Probability Model





Introduction to Logistic Regression

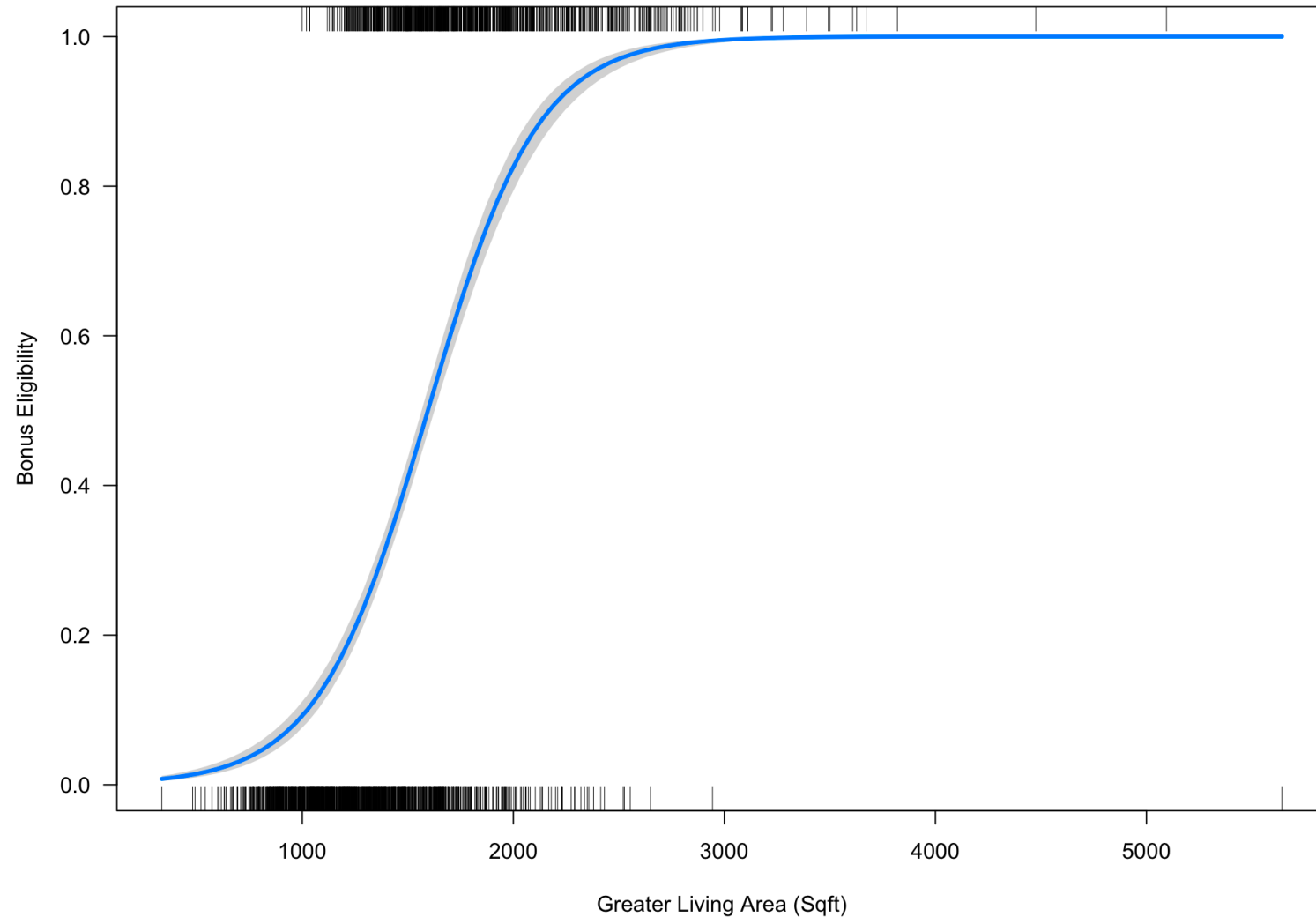
BINARY LOGISTIC REGRESSION

Logistic Regression Model

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

- Has desired properties:
 - The predicted probability will always be between 0 and 1.
 - The parameter estimates do not enter the model equation linearly.
 - The rate of change of the probability varies as the X's vary.

Logistic Regression Model



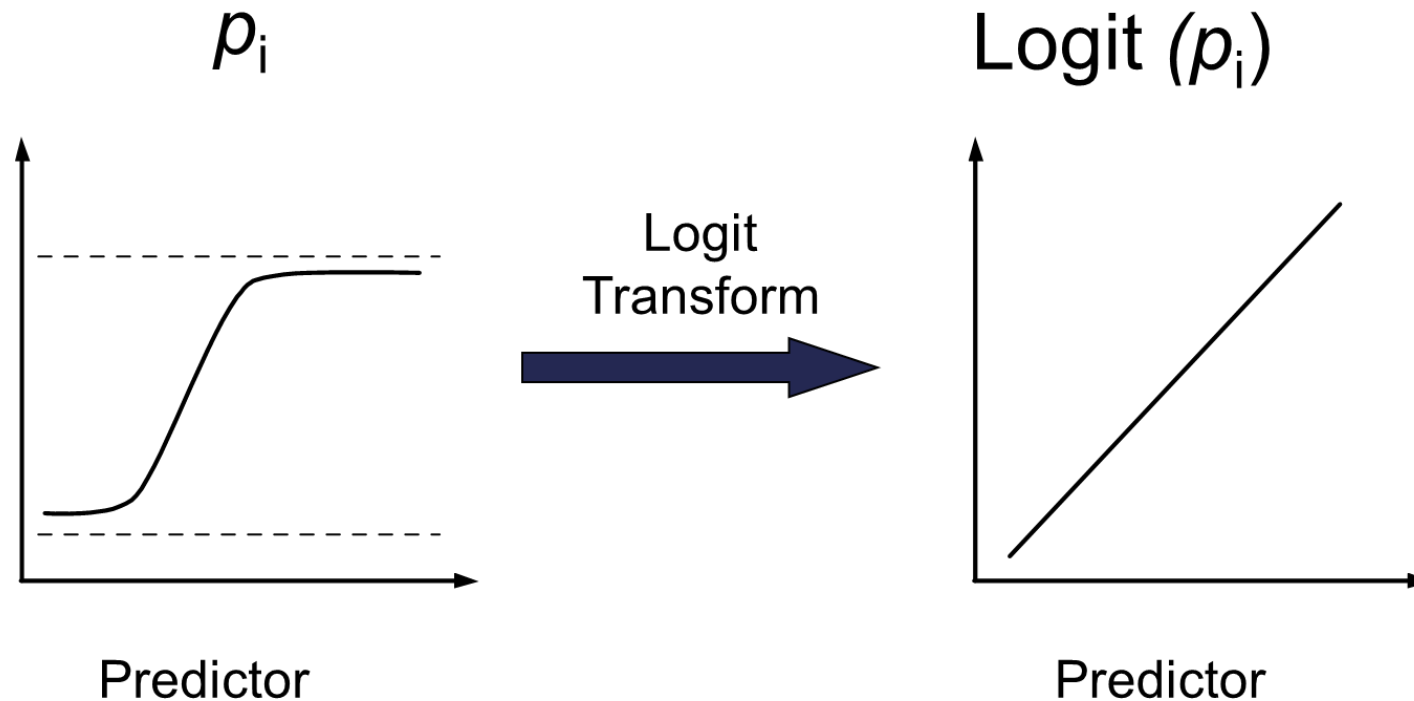
Logit Link Function

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots \beta_k x_{k,i}$$

- 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.

Assumptions

1. Independence of observations
2. Logit is linearity related to variables



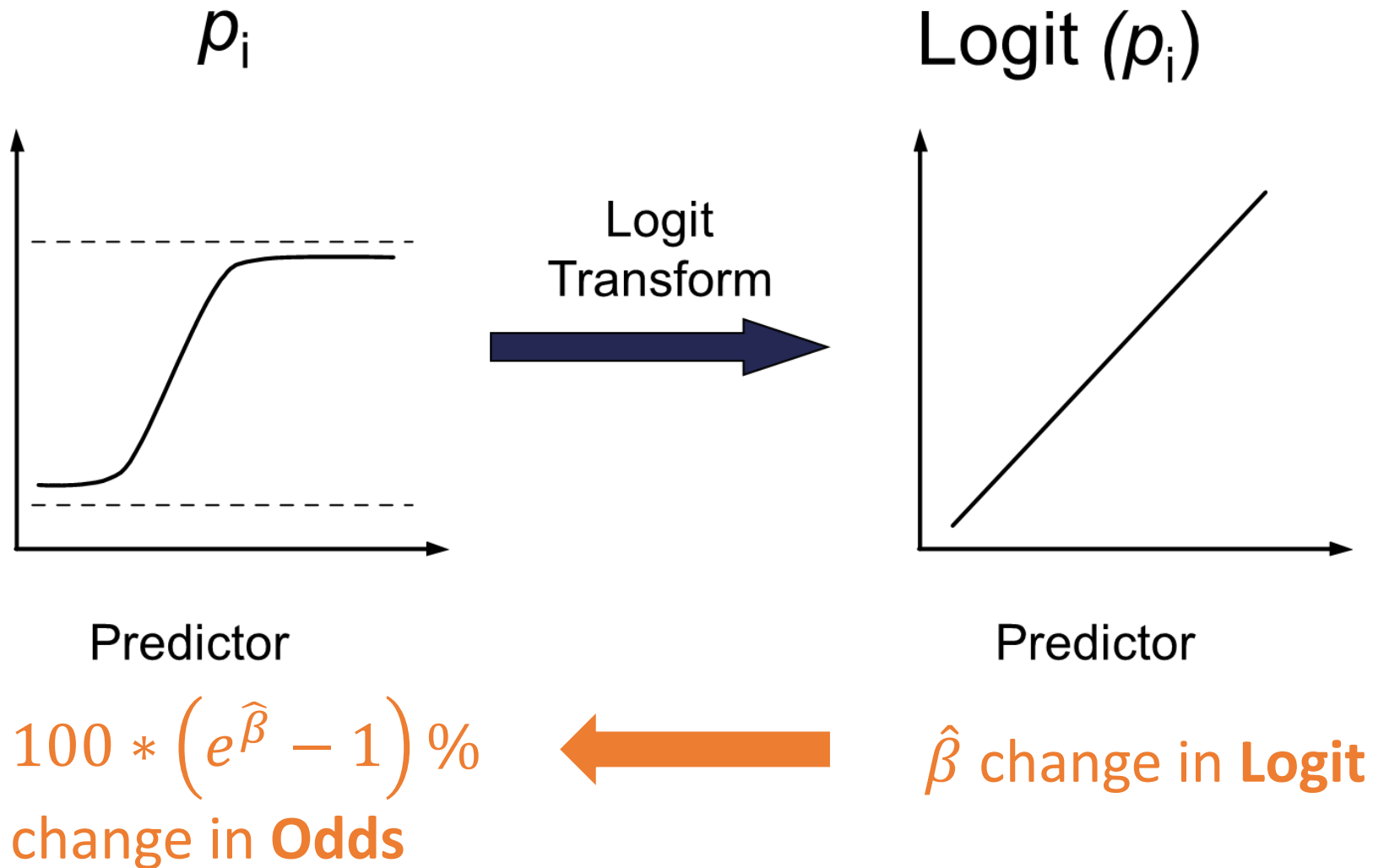
Binary Logistic Regression

```
ames_logit <- glm(Bonus ~ Gr_Liv_Area, data = train,  
                  family = binomial(link = "logit"))  
summary(ames_logit)
```


Binary Logistic Regression

```
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -5.5796  -0.6942  -0.3647   0.8060   2.1857
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.1348858  0.2757473  -22.25  <2e-16 ***
## Gr_Liv_Area  0.0038463  0.0001799   21.38  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 2775.8  on 2050  degrees of freedom
## Residual deviance: 1926.4  on 2049  degrees of freedom
## AIC: 1930.4
##
## Number of Fisher Scoring iterations: 5
```

Unit Change in Predictor does...?



Odds Ratio from a Logistic Regression

- Estimated logistic regression model:

$$\text{logit}(p_i) = -6.13 + 0.0038 * \text{Gr_Liv_Area} + \dots$$

- Estimated odds ratio (Additional SF of Greater Living Area):

$$\text{OR} = \frac{e^{-6.13 + 0.0038(GLA+1)+\dots}}{e^{-6.13 + 0.0038(GLA)+\dots}} = e^{0.0038} = 1.0038$$

$$100 \times (1.0038 - 1)\% = 0.38\%$$

- Every additional square foot of greater living area **increases the expected odds of being bonus eligible by 0.38%.**

Odds Ratios

```
100*(exp(cbind(coef(ames_logit), confint(ames_logit)))-1)
```

```
##                2.5 %      97.5 %  
## (Intercept) -99.7834027 -99.8755103 -99.6328865  
## Gr_Liv_Area  0.3853699  0.3508132  0.4216532
```

Odds Ratio from a Logistic Regression

- Estimated logistic regression model:

$$\text{logit}(p_i) = -6.13 + 0.0038 * \text{Gr_Liv_Area} + \dots$$

- Estimated odds ratio (Additional 100 SF of Greater Living Area):

$$\text{OR} = \frac{e^{-6.13 + 0.0038(\text{GLA}+100)+\dots}}{e^{-6.13 + 0.0038(\text{GLA})+\dots}} = e^{0.0038 \times 100} = 1.46$$

$$100 \times (1.46 - 1)\% = 46\%$$

- Every additional 100 square foot of greater living area **increases the expected odds of being bonus eligible by 46%.**

Odds Ratio for a Categorical Variable

- Estimated logistic regression model:

$$\text{logit}(p_i) = -9.97 + 3.56 * \text{Central_AirY} + \dots$$

- Estimated odds ratio (Central Air Y vs. N):

$$\text{OR} = \frac{e^{-9.97 + 3.56(\text{CA} \times 1) + \dots}}{e^{-9.97 + 3.56(\text{CA} \times 0) + \dots}} = e^{3.56} = 35.16$$

$$100 \times (35.16 - 1)\% = 3416\%$$

- Homes with central air **increases the expected odds of being bonus eligible by 3416%** compared to those without central air.

Odds Ratio for a Categorical Variable

- Estimated logistic regression model:

$$\text{logit}(p_i) = -9.97 + 3.56 * \text{Central_AirY} + \dots$$

- Estimated odds ratio (Central Air Y vs. N):

$$\text{OR} = \frac{e^{-9.97 + 3.56(\text{CA} \times 1) + \dots}}{e^{-9.97 + 3.56(\text{CA} \times 0) + \dots}} = e^{3.56} = 35.16$$

$$100 \times (35.16 - 1)\% = 3416\%$$

- Homes with central air **are 35.16 times as likely to be bonus eligible** then compared to those without central air.

Odds Ratios

```
ames_logit2 <- glm(Bonus ~ Gr_Liv_Area + Central_Air + factor(Fireplaces),  
                  data = train, family = binomial(link = "logit"))  
100*(exp(cbind(coef(ames_logit2), confint(ames_logit2)))-1)
```

Coefficients:

##		Estimate	Std. Error	z value	Pr(> z)	
##	(Intercept)	-9.970e+00	6.549e-01	-15.223	< 2e-16	***
##	Gr_Liv_Area	3.759e-03	2.031e-04	18.506	< 2e-16	***
##	Central_AirY	3.564e+00	5.310e-01	6.711	1.93e-11	***
##	factor(Fireplaces)1	9.822e-01	1.253e-01	7.837	4.60e-15	***
##	factor(Fireplaces)2	6.734e-01	2.406e-01	2.799	0.00513	**
##	factor(Fireplaces)3	-3.993e-02	8.711e-01	-0.046	0.96344	
##	factor(Fireplaces)4	9.025e+00	3.247e+02	0.028	0.97783	



Introduction to Logistic Regression

MODEL ASSESSMENT

Assessment of Logistic Regression

- **Many** different ways to assess logistic regression models
- One foundational way to evaluate models are comparing every pair of 0's and 1's in the target variable.
- These pairs are either considered **concordant, discordant, or tied**.

Concordant Pair

- A **concordant** pair is a 0 and 1 pair where the bonus eligible home (the 1 in our model) has a higher predicted probability than the non-bonus eligible home (the 0 in our model).
- Model successfully ordered these two observations by probability.
- It does not matter what the actual predicted probability values are as long as the bonus eligible home has a higher predicted probability than the non-bonus eligible home.

Discordant Pair

- A **discordant** pair is a 0 and 1 pair where the bonus eligible home (the 1 in our model) has a lower predicted probability than the non-bonus eligible home (the 0 in our model).
- Model unsuccessfully ordered the homes.
- It does not matter what the actual predicted probability values are as long as the bonus eligible home has a lower predicted probability than the non-bonus eligible home.

Tied Pair

- A **tied** pair is a 0 and 1 pair where the bonus eligible home has the same predicted probability as the non-bonus eligible home.
- Model is confused and sees these two different things as the same.
- In general, you want a high percentage of concordant pairs and low percentages of discordant and tied pairs.

Concordance

```
library(survival)
survival::concordance(ames_logit)
```

```
## Call:
## concordance.lm(object = ames_logit)
##
## n= 2051
## Concordance= 0.8765 se= 0.007326
## concordant discordant tied.x tied.y tied.xy
##      898616      126370      582   1074637      2070
```

Our model correctly ranks bonus eligible homes ahead of non-bonus eligible homes 87.7% of the time!

OR

Our model correctly assigns higher probability to bonus eligible homes 87.7% of the time.

Concordance

```
library(survival)
survival::concordance(ames_logit)
```

```
## Call:
## concordance.lm(object = ames_logit)
##
## n= 2051
## Concordance= 0.8765 se= 0.007326
## concordant discordant tied.x tied.y tied.xy
##      898616      126370      582   1074637      2070
```

Our model correctly ranks bonus
eligible homes ahead of non-bonus
eligible homes 87.7% of the time!

NOT

Our model is accurate 87.7% of the time.

Introduction to Logistic Regression

VARIABLE SELECTION AND REGULARIZED REGRESSION

Variable Selection

- All of the same approaches to variable selection available in linear regression are available in logistic regression.
- Forward, backward, stepwise, LASSO, etc.

Forward and Backward Selection

```
train_sel_log <- train %>%  
  dplyr::select(Bonus, Lot_Area, Street, Bldg_Type, House_Style,  
                Overall_Qual, Roof_Style, Central_Air,  
                First_Flr_SF, Second_Flr_SF, Full_Bath, Half_Bath,  
                Fireplaces, Garage_Area, Gr_Liv_Area,  
                TotRms_AbvGrd) %>%  
  mutate_if(is.numeric, ~replace_na(., mean(., na.rm = TRUE)))  
full.model <- glm(Bonus ~ . , data = train_sel_log)  
empty.model <- glm(Bonus ~ 1, data = train_sel_log)
```

Forward and Backward Selection

```
for.model <- step(empty.model,  
  scope = list(lower = formula(empty.model),  
               upper = formula(full.model)),  
  direction = "forward",  
  k = log(dim(train_sel_log)[1]))  
  
back.model <- step(full.model,  
  scope = list(lower = formula(empty.model),  
               upper = formula(full.model)),  
  direction = "backward",  
  k = log(dim(train_sel_log)[1]))
```

Results

```
## Step:  AIC=947.31
## Bonus ~ Lot_Area + Bldg_Type + Overall_Qual + First_Flr_SF +
##      Full_Bath + Half_Bath + Fireplaces + Garage_Area
```

```
## Call:
## concordance.lm(object = ames_logit2)
##
## n= 2051
## Concordance= 0.8953 se= 0.006677
## concordant discordant   tied.x   tied.y   tied.xy
##      918066      107238       264  1075521       1186
```

Regularized Regression

- Although not shown here, regularized regression can use the same link function to obtain logistic regression
- Ridge, LASSO, Elastic Net

