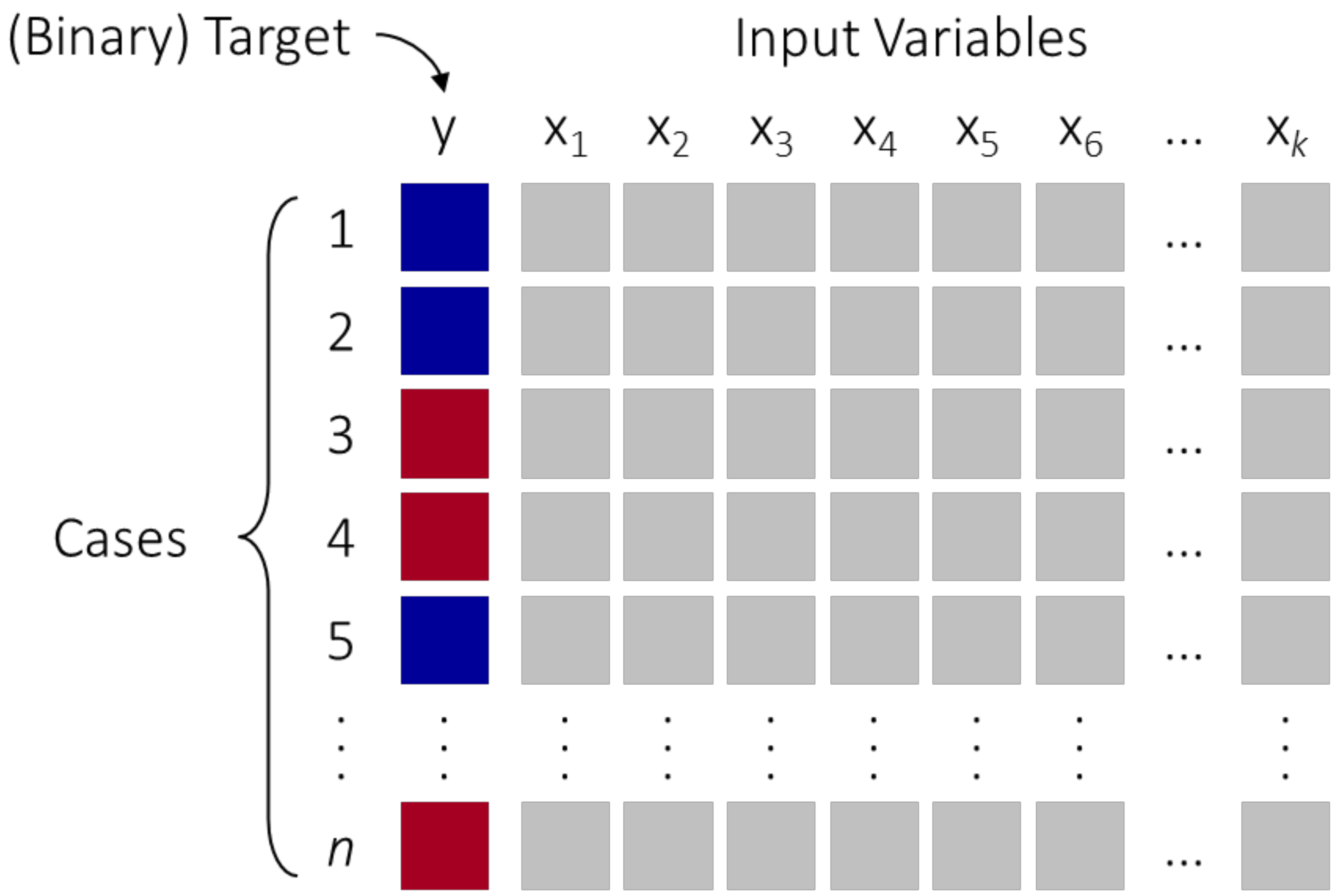
BINARY LOGISTIC REGRESSION

Dr. Aric LaBarr

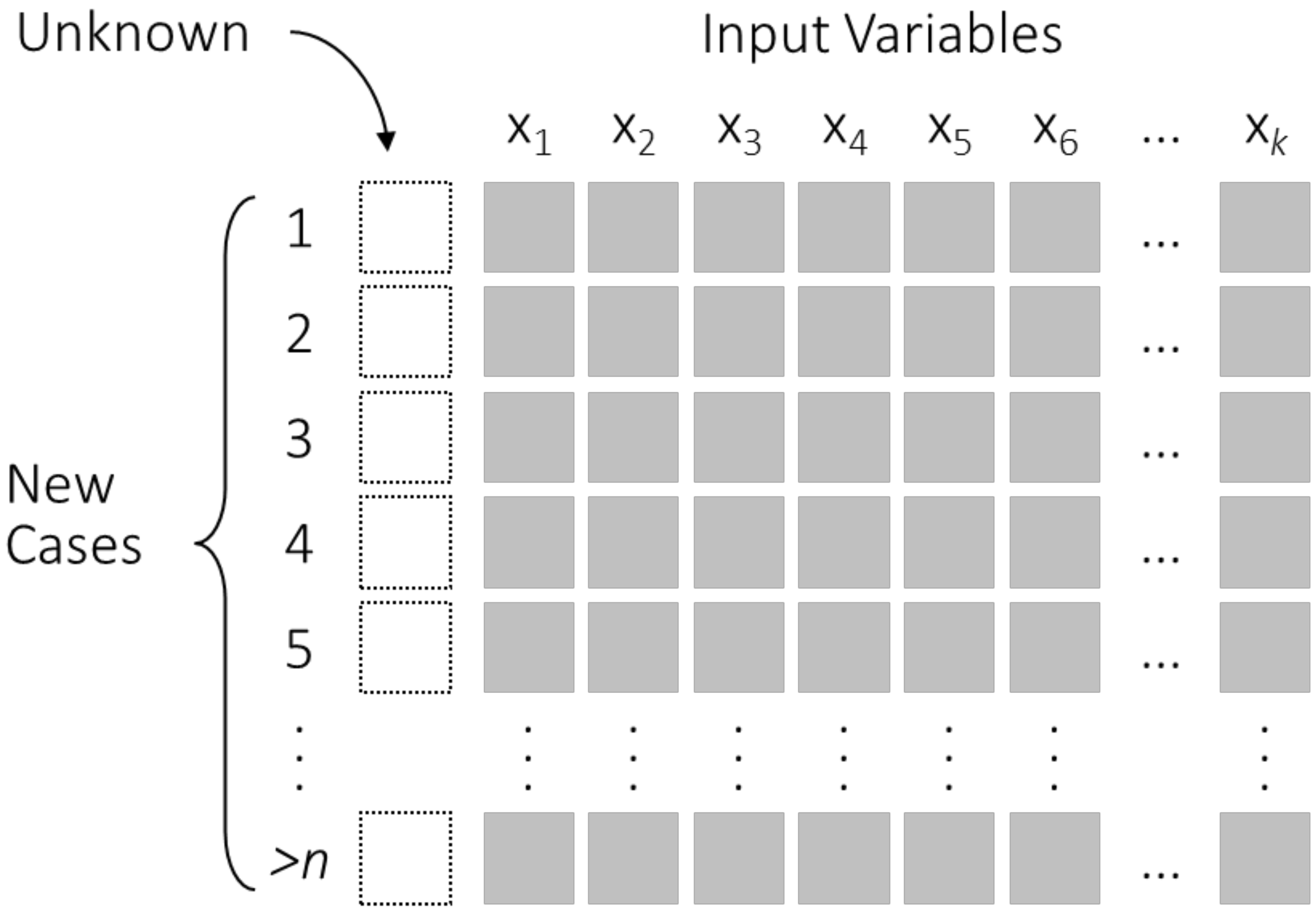
Institute for Advanced Analytics

BINARY LOGISTIC REGRESSION

Supervised Classification Modeling



Unsupervised Classification Scoring



Applications

- Binary classification is one of, if not the, **most common** type of business problems that need solving.

- Models developed by alumni in **current** jobs:
  - Targeted Marketing
  - Churn Prediction
  - Probability of Default
  - Fraud Detection

Ames Real Estate Data

- 2930 homes in Ames, Iowa in the early 2000's.
- Physical attributes of homes along with sales price of home.

Bonus Eligibility

```r
library(AmesHousing)

ames <- make_ordinal_ames()

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

What is Regression Actually Doing?

- Regression is modeling the **expected** (mean/average) response conditional on the predictors → $E(y_i|x_1, x_2, \ldots)$
- For a binary (0/1) response y_i, the expected value is just the probability of the event:

$$E(y_i) = P(y_i = 1) = p_i$$

- So why not model the following:

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

Linear Probability Model

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

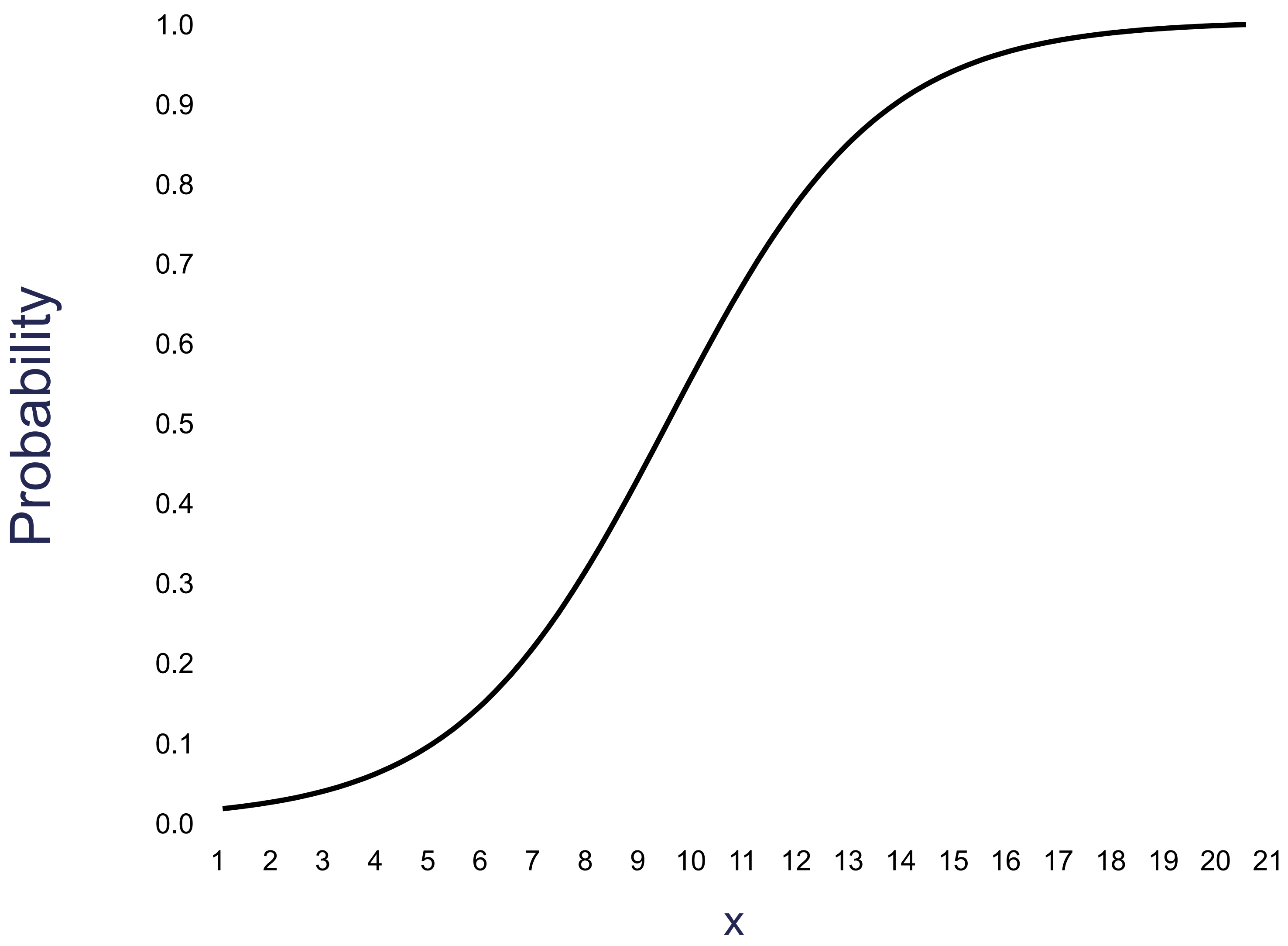
- Problems:
  - Probabilities are bounded, but linear functions can take on any value. (How do you interpret a predicted value of -0.4 or 1.1?)
  - The relationship between probabilities and X is usually nonlinear. Example, one unit change in X will have different effects when the probability is near 1 or 0.5.
  - Properties of OLS do not hold.

Logistic Regression Model

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

- 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 Curve

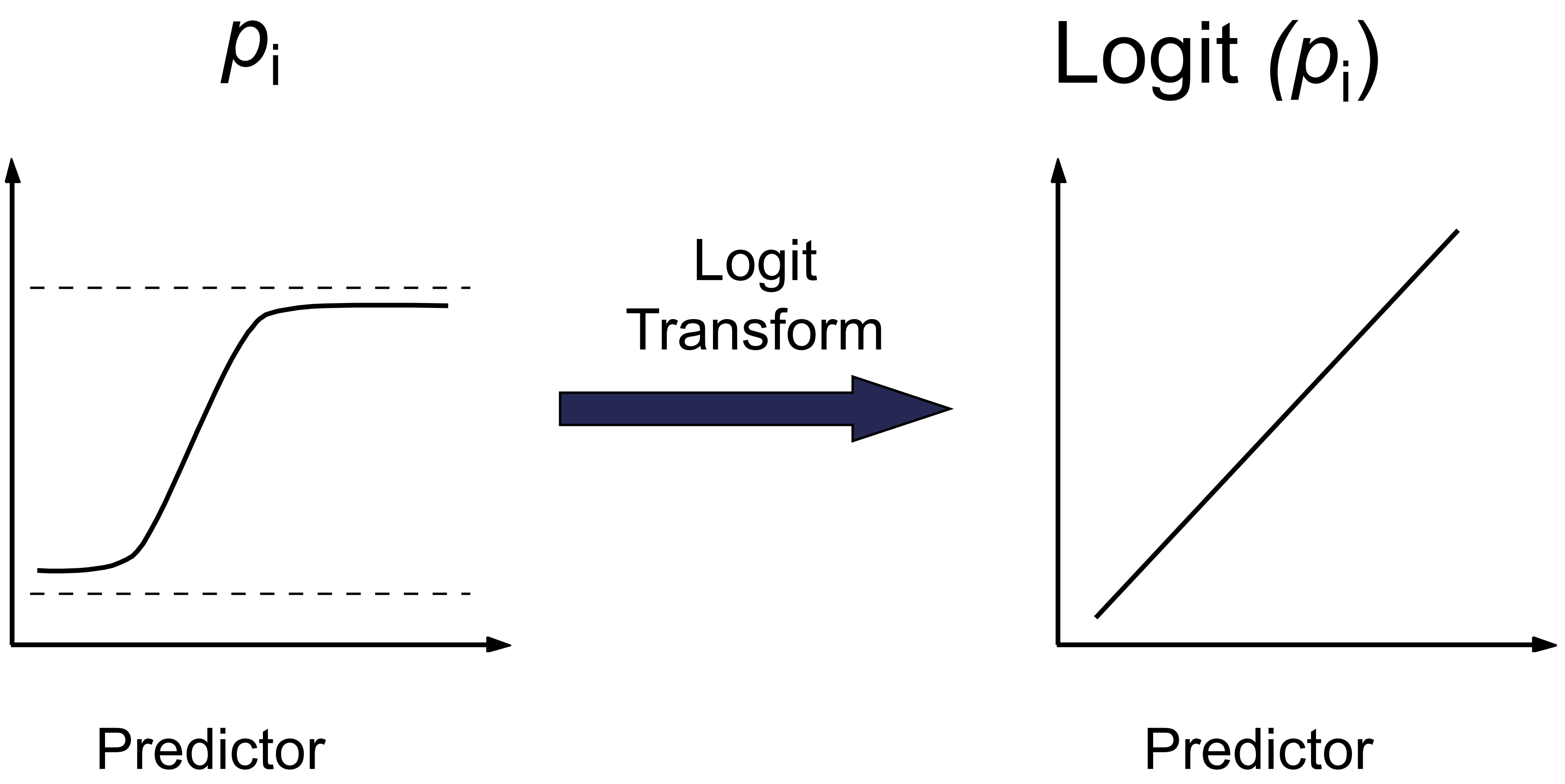


The Logit Link Transformation

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

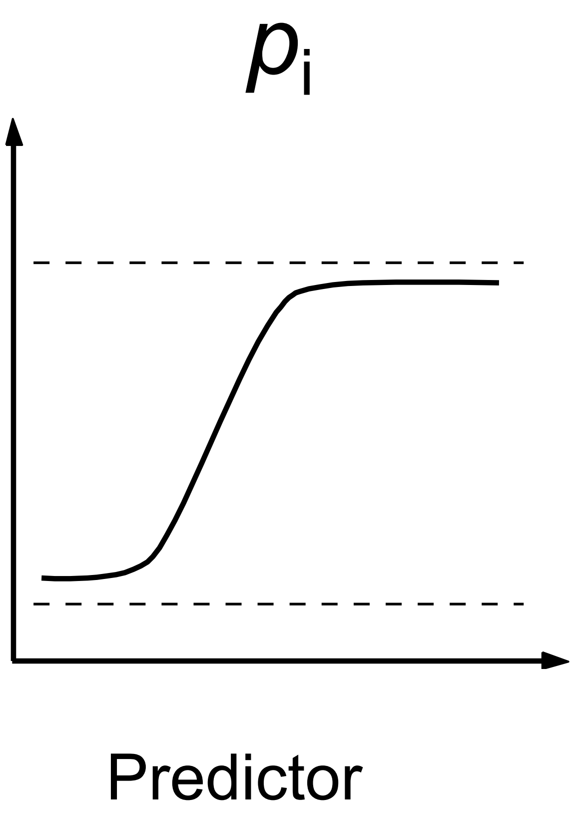
- To create a linear model, a link function (logit) is applied to the probabilities.
- The relationship between the parameters and the logits are linear.
- Logits unbounded.

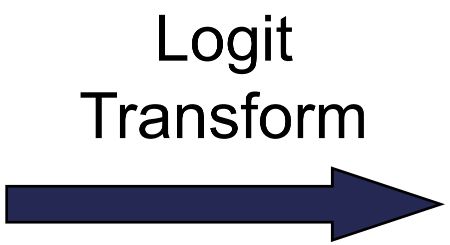
The Logit Link Transformation

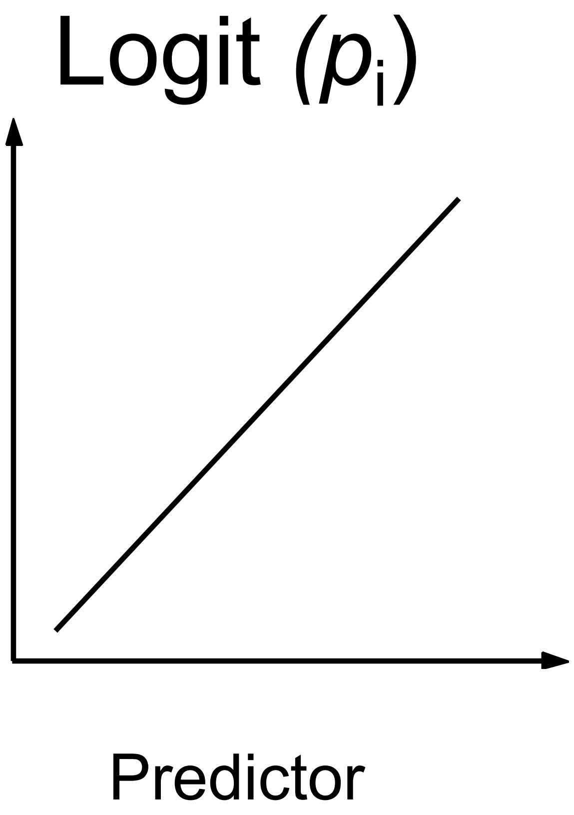


COEFFICIENT INTERPRETATIONS

Unit Change in Predictor does…?







$100 * \left(e^{\widehat{\beta}} - 1\right)\%$ change in **Odds**

$\hat{\beta}$ change in **Logit**

Estimating Coefficients

```r
logit.model <- glm(Bonus ~ Gr_Liv_Area + factor(Central_Air),
                   data = train, family = binomial(link = "logit"))

summary(logit.model)
```

Estimating Coefficients

```
Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-5.7966  -0.6628  -0.3223   0.7331   2.8308  

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -1.035e+01  6.422e-01  -16.12  < 2e-16 ***
Gr_Liv_Area          4.112e-03  1.962e-04   20.96  < 2e-16 ***
factor(Central_Air)Y  3.952e+00  5.180e-01    7.63 2.35e-14 ***
---
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: 1808.8  on 2048  degrees of freedom
AIC: 1814.8
```

Odds Ratio from a Logistic Regression

- Estimated logistic regression model:

$$\text{logit}(p_i) = -10.35 + 3.952 * Central\_AirY + \cdots$$

- Estimated odds ratio (Central Air vs. No Central Air):

$$\text{OR} = \frac{e^{-10.35+3.952(1)+\cdots}}{e^{-10.35+3.952(0)+\cdots}} = \frac{e^{-10.35}e^{3.952}\ \ldots}{e^{-10.35}\ \ldots} = e^{3.952} = 52.03$$

- Homes with central air have $\mathbf{100 * (e^{3.952} - 1)\% = 5103\%}$ **higher expected odds** than homes without central air to be bonus eligible.

Odds Ratio from a Logistic Regression

- Estimated logistic regression model:

$$\text{logit}(p_i) = -10.35 + 3.952 * Central\_AirY + \cdots$$

- Estimated odds ratio (Central Air vs. No Central Air):

$$\text{OR} = \frac{e^{-10.35+3.952(1)+\cdots}}{e^{-10.35+3.952(0)+\cdots}} = \frac{e^{-10.35}e^{3.952}\ \ldots}{e^{-10.35}\ \ldots} = e^{3.952} = 52.03$$

- Homes with central air are **52.03 times more likely** to be bonus eligible than homes without central air, on average.

Odds Ratio from a Logistic Regression

- Estimated logistic regression model:

$$\text{logit}(p_i) = -10.35 + 0.0041 * Gr\_Liv\_Area + \cdots$$

- Estimated odds ratio (Additional Square Foot of Space):

$$\text{OR} = \frac{e^{-10.35+0.0041(Gr\_Liv\_Area+1)+\cdots}}{e^{-10.35+0.0041(Gr\_Liv\_Area)+\cdots}} = \frac{e^{-10.35}e^{0.0041}\ \ldots}{e^{-10.35}\ \ldots} = e^{0.0041} = 1.0041$$

- Every additional square foot of space expects to have $\mathbf{100 * \left(e^{0.0041} - 1\right)\% = 0.41\%}$ **higher odds** to be bonus eligible.

Amount to Double the Odds

- Working through the math backwards allows us to see what increase in square footage is needed for an expected doubling of the odds of a home being bonus eligible.
- Estimated logistic regression model:

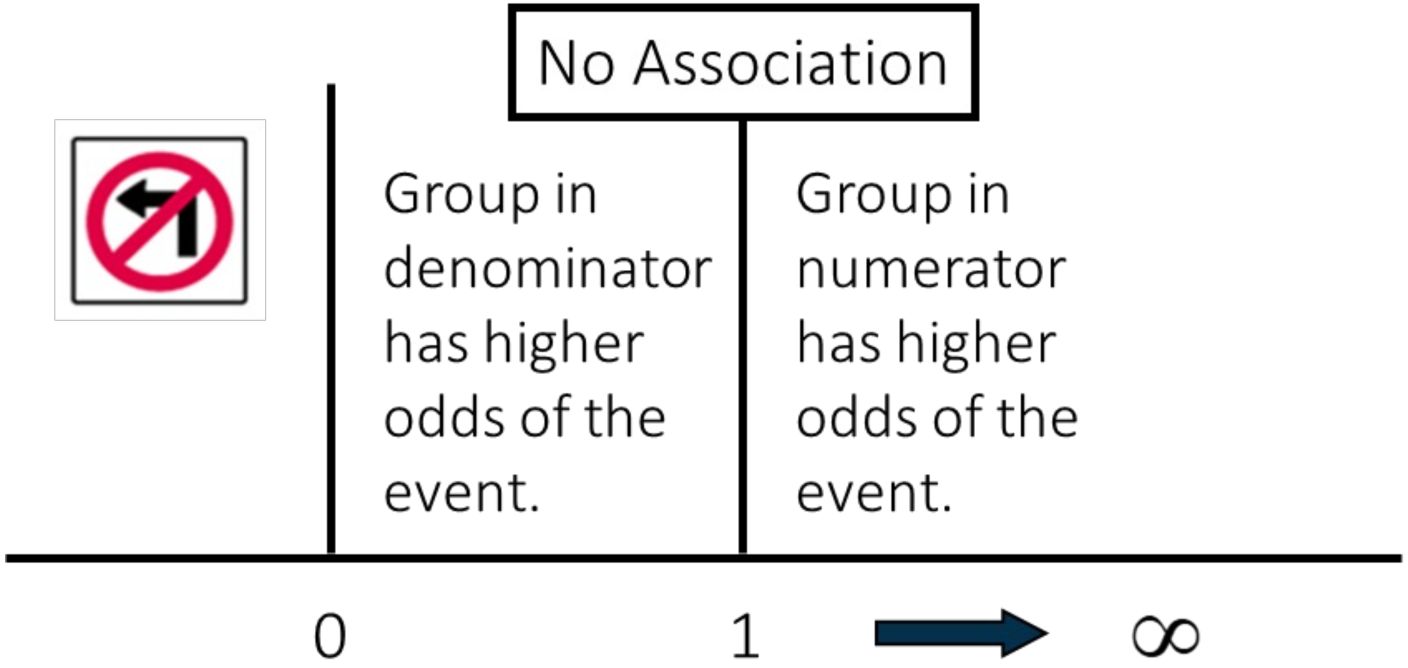
$$\operatorname{logit}(p_i) = -10.35 + 0.0041 * Gr\_Liv\_Area + \cdots$$

- Estimated amount to double the odds:

$$Double\ Odds = \frac{\log(2)}{\beta} = \frac{\log(2)}{0.0041} = 169.06$$

- Every additional square foot of space increase of 169.06 **doubles the odds** to be bonus eligible.

Properties of the Odds Ratio



Odds Ratio

```
exp(
    cbind(coef(logit.model), confint(logit.model))
    )
```

```
                                       2.5 %       97.5 %
(Intercept)          3.184558e-05 8.233966e-06 1.041852e-04
Gr_Liv_Area          1.004121e+00 1.003745e+00 1.004517e+00
factor(Central_Air)Y 5.203450e+01 2.058035e+01 1.620722e+02
```

ESTIMATION METHOD

Assumptions for OLS Regression

- The random error term has a Normal distribution with a mean of zero.
- The random error term has constant variance.
- The error terms are independent.
- Linearity of the mean.
- No perfect collinearity.

- In logistic regression, the first two assumptions are violated. Therefore, OLS is not the best method for parameter estimation.

Maximum Likelihood Estimation

- In logistic regression, estimates are obtained via **maximum likelihood estimation (MLE)**
- Very popular technique for developed statistical models!
- In fact, OLS is mathematically the same as the maximum likelihood by (INSERT MATH HERE!)

- The **likelihood function** measures how probable a specific grid of β values is to have produced your data → so we want to MAXIMIZE that!

Likelihood Function

- The **likelihood function** measures how probable a specific grid of β values is to have produced your data → so we want to MAXIMIZE that!
- Based off the probability density function.

- Binomial target variable:

$$L(\beta' s | y, x_1, x_2, \dots) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$$

Likelihood Function

- The **likelihood function** measures how probable a specific grid of β values is to have produced your data → so we want to MAXIMIZE that!
- Based off the probability density function.

- Binomial target variable:

The 1's and their probability

$$L(\beta's|y, x_1, x_2, \dots) = \prod_{i=1}^{n} \boxed{p_i^{y_i}}(1-p_i)^{1-y_i}$$

Likelihood Function

- The **likelihood function** measures how probable a specific grid of β values is to have produced your data → so we want to MAXIMIZE that!
- Based off the probability density function.

- Binomial target variable:

The 1's and their probability

$$L(\beta's|y, x_1, x_2, \dots) = \prod_{i=1}^{n} \boxed{p_i^{y_i}} \boxed{(1-p_i)^{1-y_i}}$$

The 0's and their probability

Likelihood Function

- The **likelihood function** measures how probable a specific grid of β values is to have produced your data → so we want to MAXIMIZE that!
- Based off the probability density function.

- Binomial target variable **with logistic regression**:

$$L(\beta's|y, x_1, x_2, \dots) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$$

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

Log-Likelihood Function

- Usually easier to mathematically work with the log of the likelihood function instead.

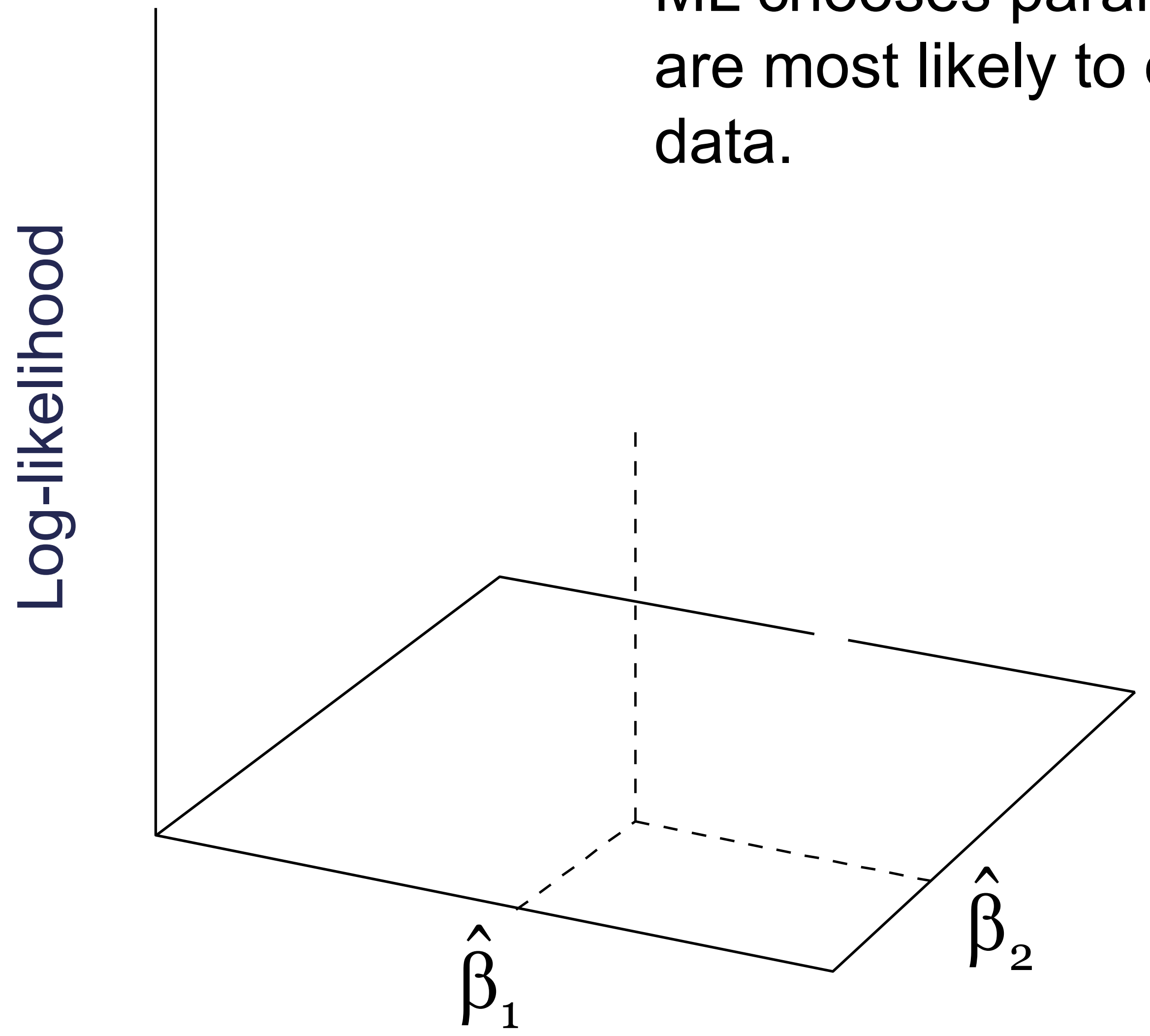
- Binomial target variable **with logistic regression**:

$$\log(L) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

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

Maximum Likelihood Estimation

ML chooses parameters that are most likely to occur, given data.

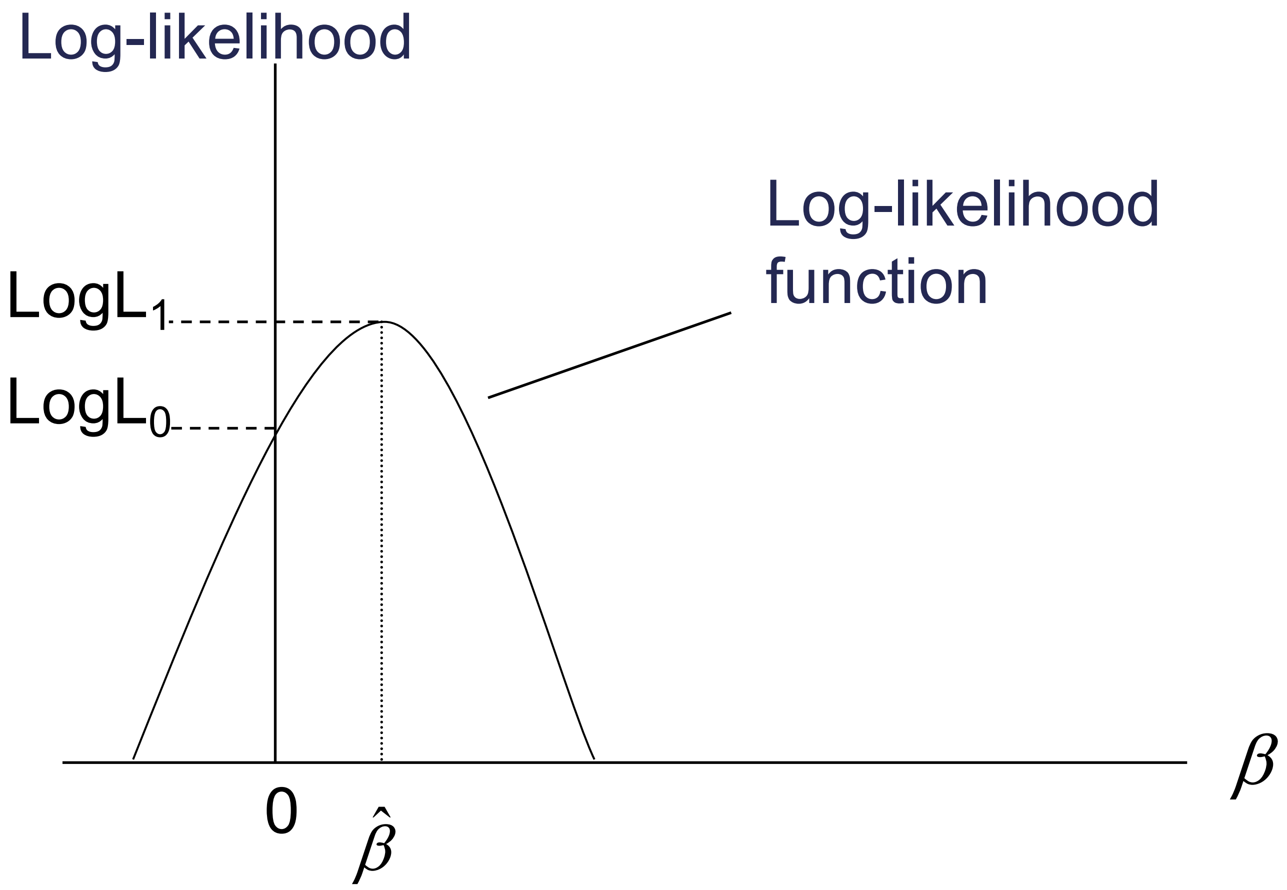


Likelihood Ratio Tests

- Likelihood estimation provides a basis for **hypothesis testing.**
- If extra predictors don't add much information, then a model that includes them shouldn't be substantially more likely than the model that doesn't include them.

- **Likelihood Ratio Test (LRT)** compares these FULL and REDUCED models.

Model Inference – Likelihood Ratio Test



LRT= -2 (LogL_0 – LogL_1), follows chi-square distribution

Likelihood Ratio Test

```
logit.model.r <- glm(Bonus ~ 1,
                     data = train, family = binomial(link = "logit"))

anova(logit.model, logit.model.r, test = 'LRT')
```

Likelihood Ratio Test

```
Analysis of Deviance Table

Model 1: Bonus ~ Gr_Liv_Area + factor(Central_Air)
Model 2: Bonus ~ 1
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      2048     1808.8
2      2050     2775.8 -2  -966.96  < 2.2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

LRT Used for Categorical P-values

- Shouldn't use design variable p-values for categorical variables with more than 2 levels → NOT ALL COMPARISONS ARE SHOWN!
- Use Likelihood Ratio Test to compare model with and without the categorical variable.

- If different (low p-value), then model with categorical variable provides additional information.
- If not different (high p-value), then model with categorical variable doesn't provide additional information (can drop variable).

Likelihood Ratio Test

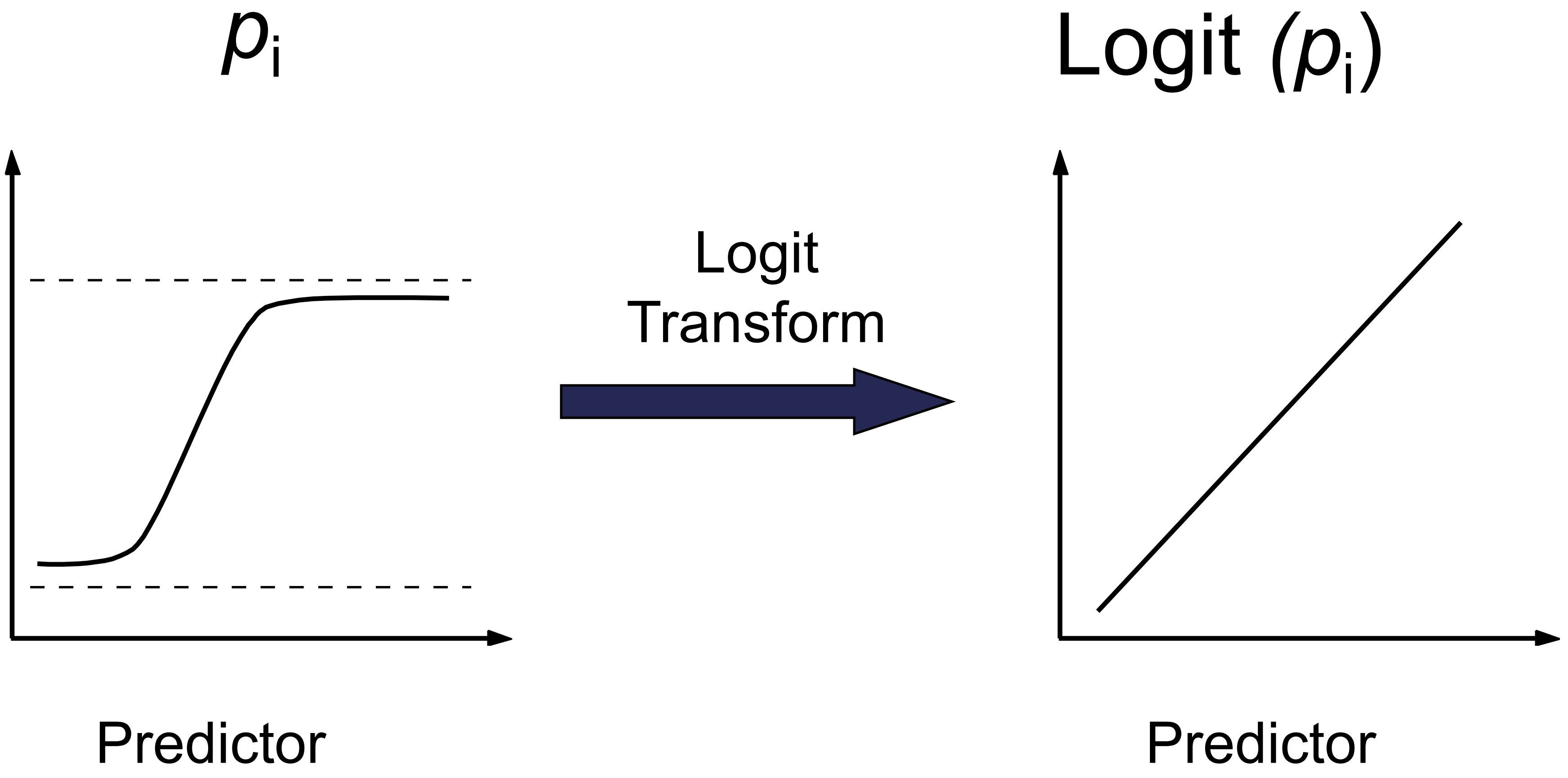
```
logit.model.f <- glm(Bonus ~ Gr_Liv_Area +
                             factor(Central_Air) +
                             factor(Fireplaces),
                     data = train, family = binomial(link = "logit"))

car::Anova(logit.model.f, test = 'LR', type = 'III')
```

Likelihood Ratio Test

```
Analysis of Deviance Table (Type III tests)

Response: Bonus

                   LR Chisq Df Pr(>Chisq)
Gr_Liv_Area          565.89  1  < 2.2e-16 ***
factor(Central_Air)   86.81  1  < 2.2e-16 ***
factor(Fireplaces)    62.61  4  8.181e-13 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ASSUMPTIONS

Assumption



How do we evaluate?

General Additive Model (GAM)

- Traditional logistic regression model:

$$\log(odds) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}$$

- GAM logistic regression model:

$$\log(odds) = \beta_0 + f_1\left(x_{1,i}\right) + \cdots + f_k\left(x_{k,i}\right)$$

- Use **spline functions** to estimate $f_j\left(x_j\right)$.
- If splines say straight line is good, then assumption met!

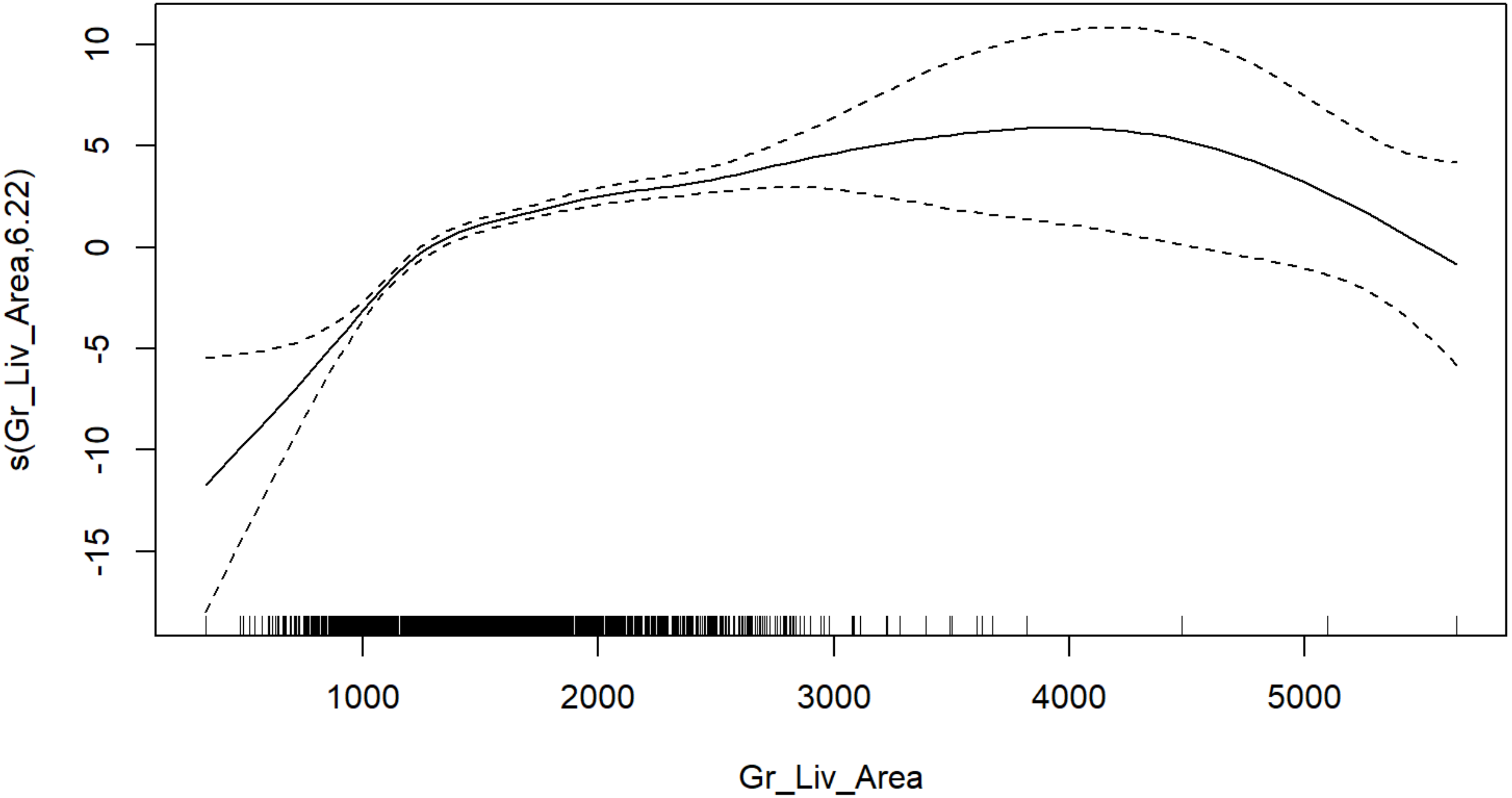
Checking Assumptions – GAM

```
fit.gam <- gam(Bonus ~ s(Gr_Liv_Area) + factor(Central_Air),
               data = train, family = binomial(link = 'logit'), method = 'REML')

summary(fit.gam)

plot(fit.gam)
```

Checking Assumptions – GAM

```
Family: binomial
Link function: logit

Formula:
Bonus ~ s(Gr_Liv_Area) + factor(Central_Air)

Parametric coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -4.4616     0.5033  -8.864  < 2e-16 ***
factor(Central_Air)Y   3.4882     0.4911   7.103 1.22e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                 edf Ref.df Chi.sq p-value
s(Gr_Liv_Area) 6.221  7.232  380.4  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =   0.43   Deviance explained =   39%
-REML = 859.46  Scale est. = 1         n = 2051
```

Checking Assumptions – GAM



Checking Assumptions – GAM

```
Family: binomial
Link function: logit

Formula:
Bonus ~ s(Gr_Liv_Area) + factor(Central_Air)

Parametric coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -4.4616     0.5033  -8.864  < 2e-16 ***
factor(Central_Air)Y   3.4882     0.4911   7.103 1.22e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                 edf Ref.df Chi.sq p-value
s(Gr_Liv_Area) 6.221  7.232  380.4  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =   0.43   Deviance explained =   39%
-REML = 859.46  Scale est. = 1         n = 2051
```

Does Spline Add Value

```
anova(logit.model, fit.gam, test = 'Chisq')
```

```
Analysis of Deviance Table

Model 1: Bonus ~ Gr_Liv_Area + factor(Central_Air)
Model 2: Bonus ~ s(Gr_Liv_Area) + factor(Central_Air)
  Resid. Df Resid. Dev     Df Deviance  Pr(>Chi)
1    2048.0     1808.8
2    2042.8     1692.3 5.2212   116.58 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

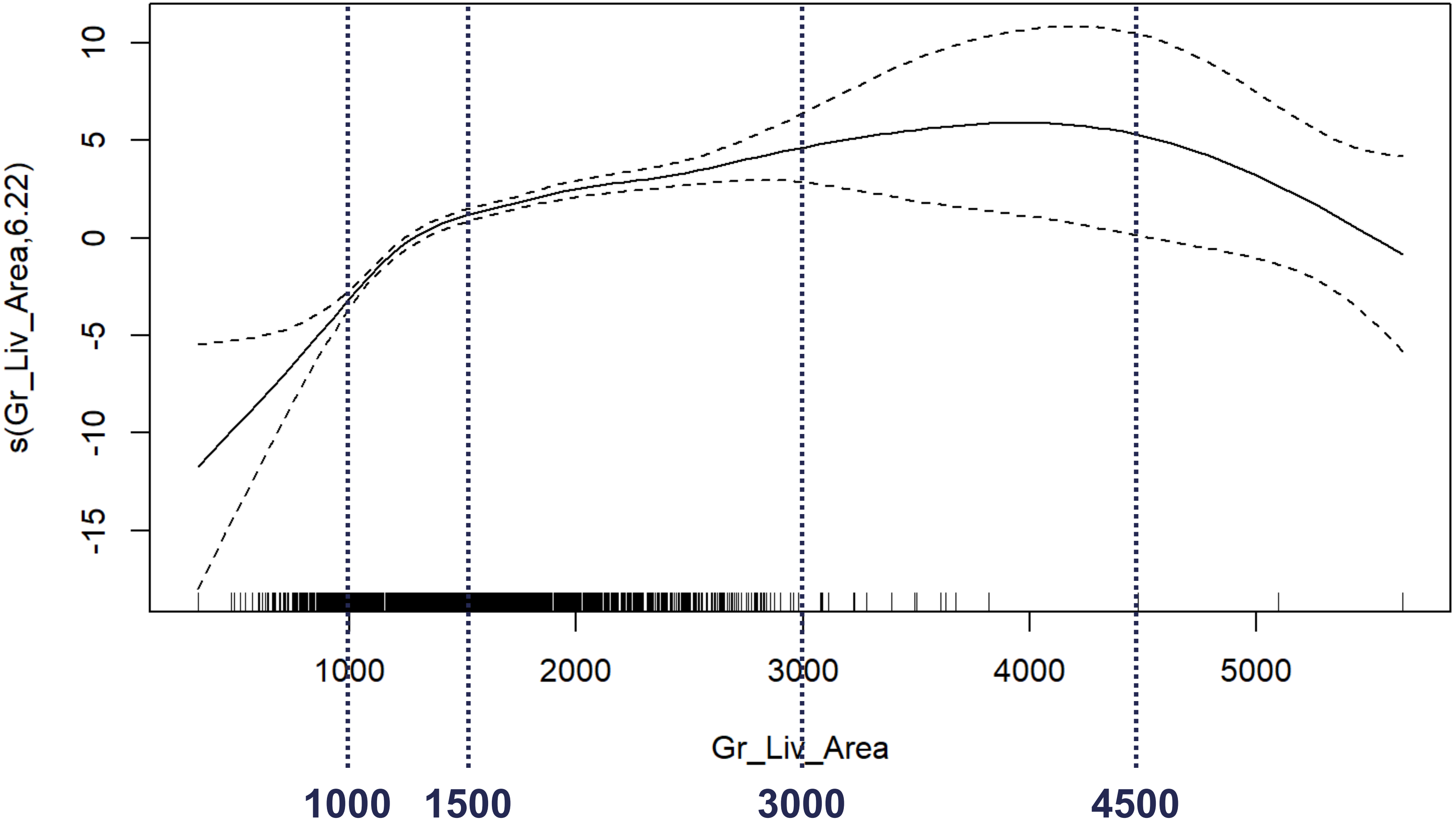
Assumptions Failed?

- What if the linearity assumption failed for at least one of the continuous variables?
  1. Use GAM logistic model instead with more limited interpretation on variables that break assumption
  2. Bin the continuous variables that break assumption (keeps interpretation ability)

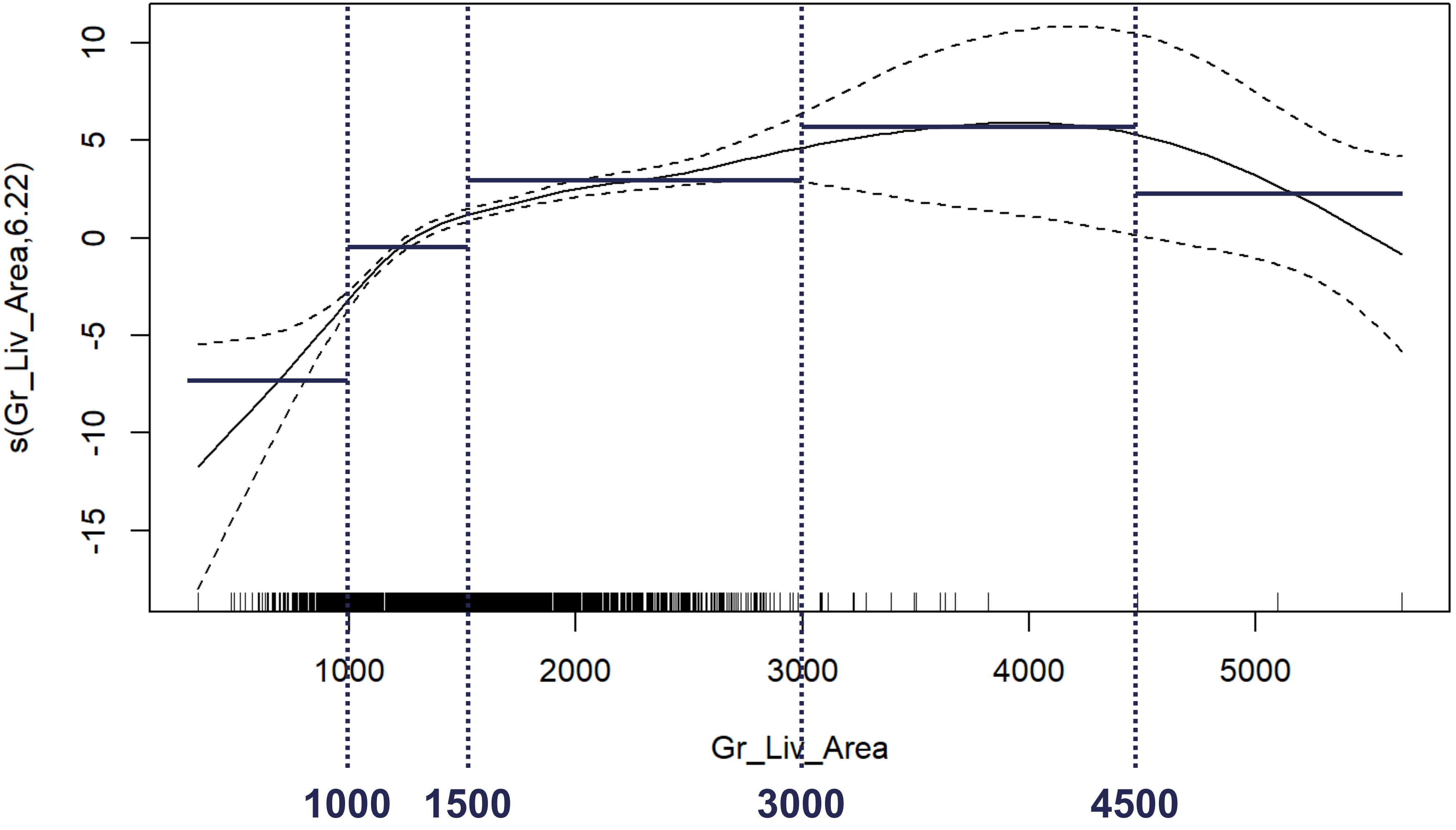
Assumptions Failed?

- What if the linearity assumption failed for at least one of the continuous variables?
  1. Use GAM logistic model instead with more limited interpretation on variables that break assumption
  2. Bin the continuous variables that break assumption (keeps interpretation ability)

Binning Continuous Variable



Binning Continuous Variable



Binning Continuous Variable

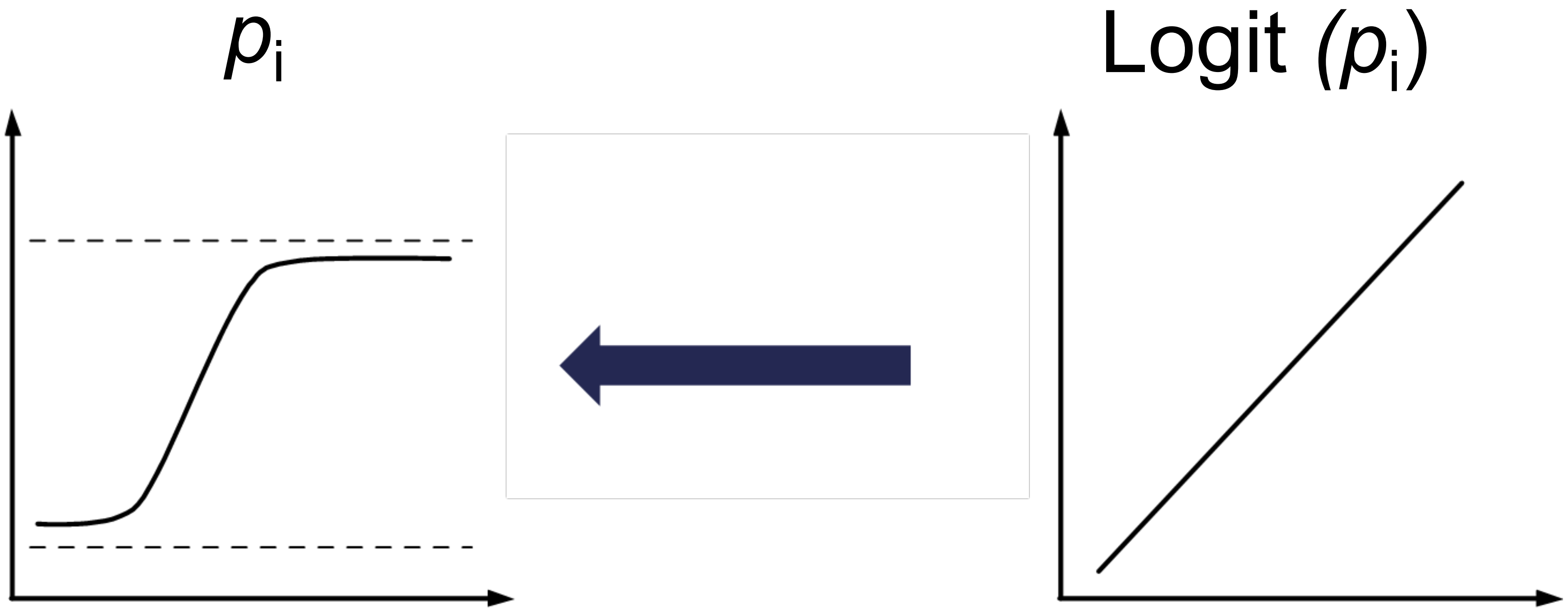
```
train <- train %>%
  mutate(Gr_Liv_Area_BIN = cut(Gr_Liv_Area,
         breaks = c(-Inf,1000,1500,3000,4500,Inf)))

logit.model.bin <- glm(Bonus ~ factor(Gr_Liv_Area_BIN) + factor(Central_Air),
                       data = train, family = binomial(link = 'logit'))

summary(logit.model.bin)
```

Binning Continuous Variable

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6410  -0.7626  -0.0860   0.7763   3.3473

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                             -8.8210     1.1065  -7.972 1.56e-15 ***
factor(Gr_Liv_Area_BIN)(1e+03,1.5e+03]   4.5121     1.0052   4.489 7.16e-06 ***
factor(Gr_Liv_Area_BIN)(1.5e+03,3e+03]   6.6437     1.0049   6.611 3.81e-11 ***
factor(Gr_Liv_Area_BIN)(3e+03,4.5e+03]  21.1646   363.8508   0.058  0.95361
factor(Gr_Liv_Area_BIN)(4.5e+03, Inf]    5.5986     1.7331   3.230  0.00124 **
factor(Central_Air)Y                     3.2224     0.4734   6.807 9.95e-12 ***
---
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: 1892.0  on 2045  degrees of freedom
AIC: 1904
```

PREDICTED VALUES

Predicted Probabilities



- Once model fitting is over, we want to convert back to probabilities for our predictions.

Predicted Values

```
new_ames <- data.frame(Gr_Liv_Area = c(1500, 2000, 2250, 2500, 3500),
                       Central_Air = c("N", "Y", "Y", "N", "Y"))

new_ames <- data.frame(new_ames,
                       'Pred' = predict(logit.model, newdata = new_ames,
                                        type = "response"))

print(new_ames)
```

Predicted Values

```
 Gr_Liv_Area Central_Air       Pred
1        1500           N 0.01498152
2        2000           Y 0.86084436
3        2250           Y 0.94534188
4        2500           N 0.48167577
5        3500           Y 0.99966165
```

Predicted Probability Plot – R