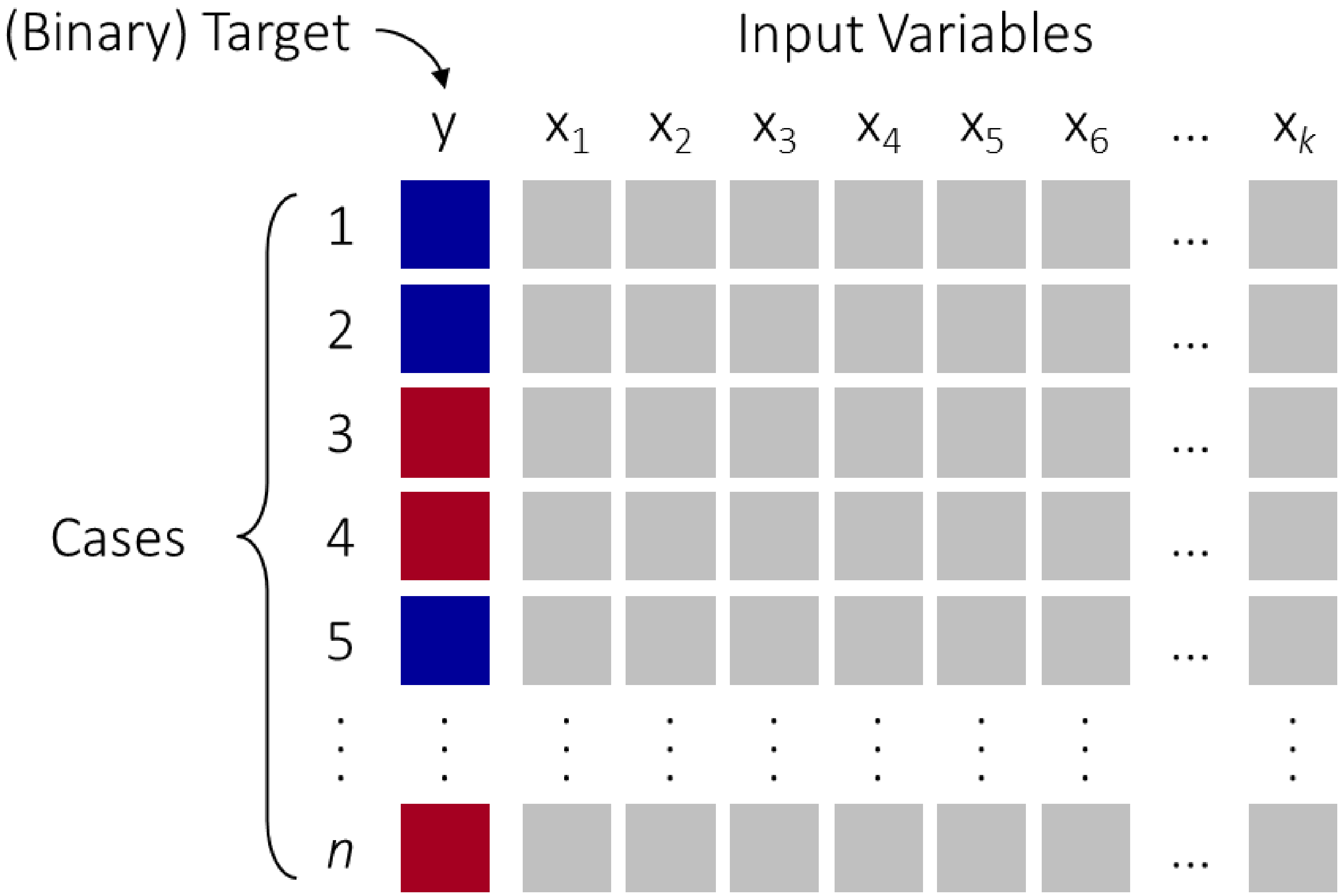
BINARY LOGISTIC REGRESSION

Dr. Aric LaBarr

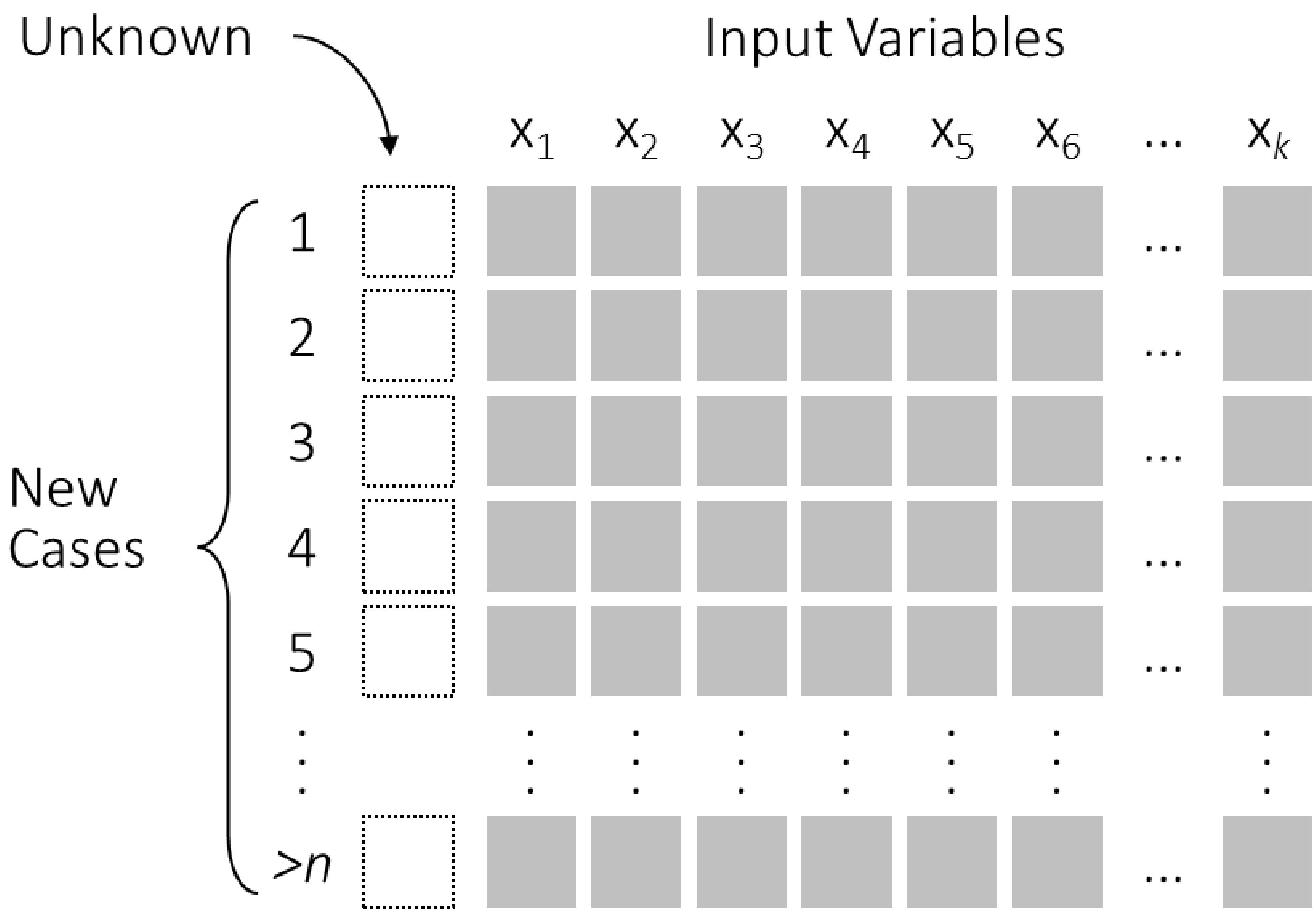
Institute for Advanced Analytics

BINARY LOGISTIC REGRESSION

Supervised Classification Modeling



Unsupervised Classification Scoring



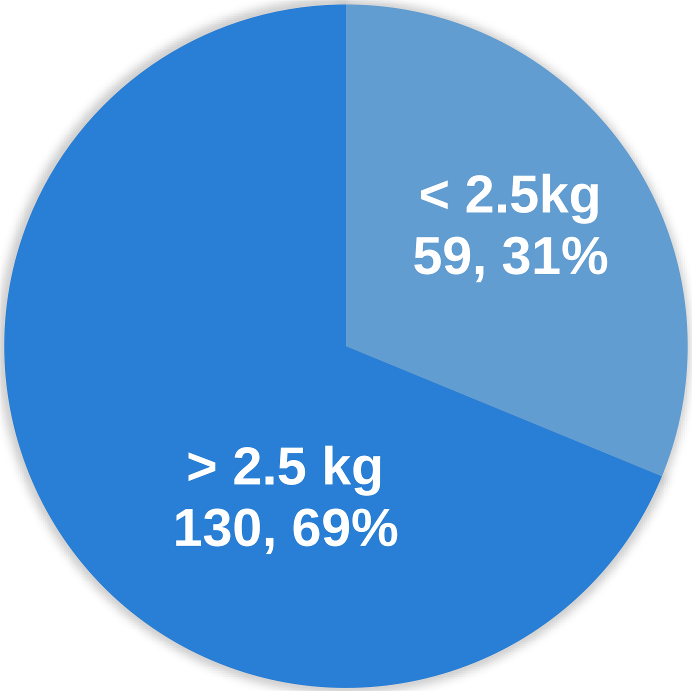
Applications

- 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

Birth Weight Data Set

- Model the association between various factors and child being born with low birth weight (< 2.5kg)
- 189 observations in the data set



Birth Weight Data Set

- Model the association between various factors and child being born with low birth weight (< 2.5kg)
- Predictors:
  - **age:** mother's age (years)
  - **lwt:** mother's weight at last menstrual period (lbs)
  - **smoke:** mother's smoking status during pregnancy
  - **race:** mother's race (1=White, 2 = Black, 3 = Other)
  - **ptl:** number of premature labors
  - **ht:** history of hypertension
  - **ui:** uterine irritability
  - **ftv:** number of physician visits during first trimester

What is Regression Actually Doing?

- Regression is modeling the **expected** (mean/average) response conditional on the predictors → $E(y_i | x_1, x_2, \dots)$
- 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

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

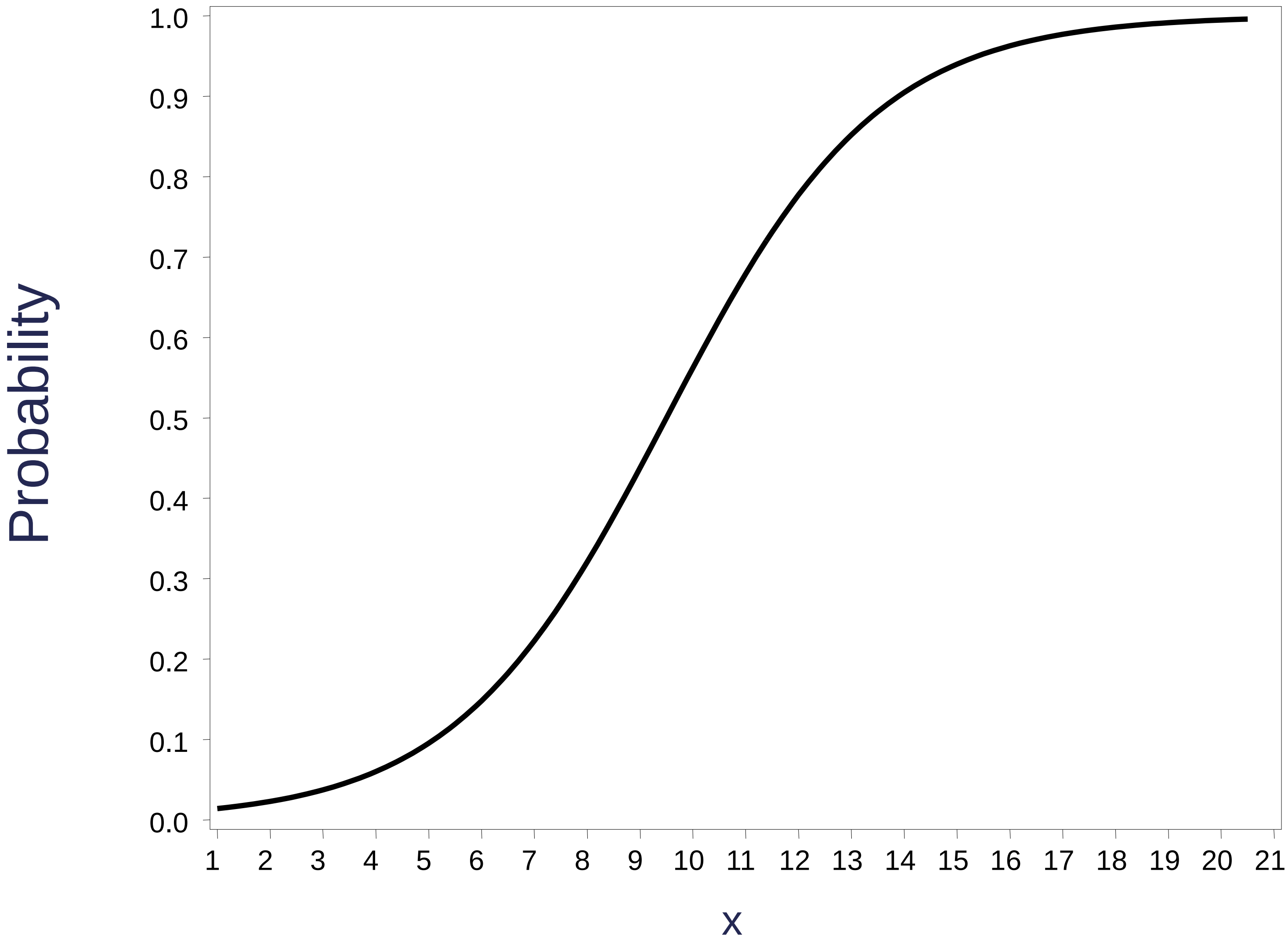
- Problems:
  - Probabilities are bounded, but linear functions can take on any value. (How do you interpret a predicted value of -0.4 or 1.1?)
  - The relationship between probabilities and X is usually nonlinear. Example, one unit change in X will have different effects when the probability is near 1 or 0.5.

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

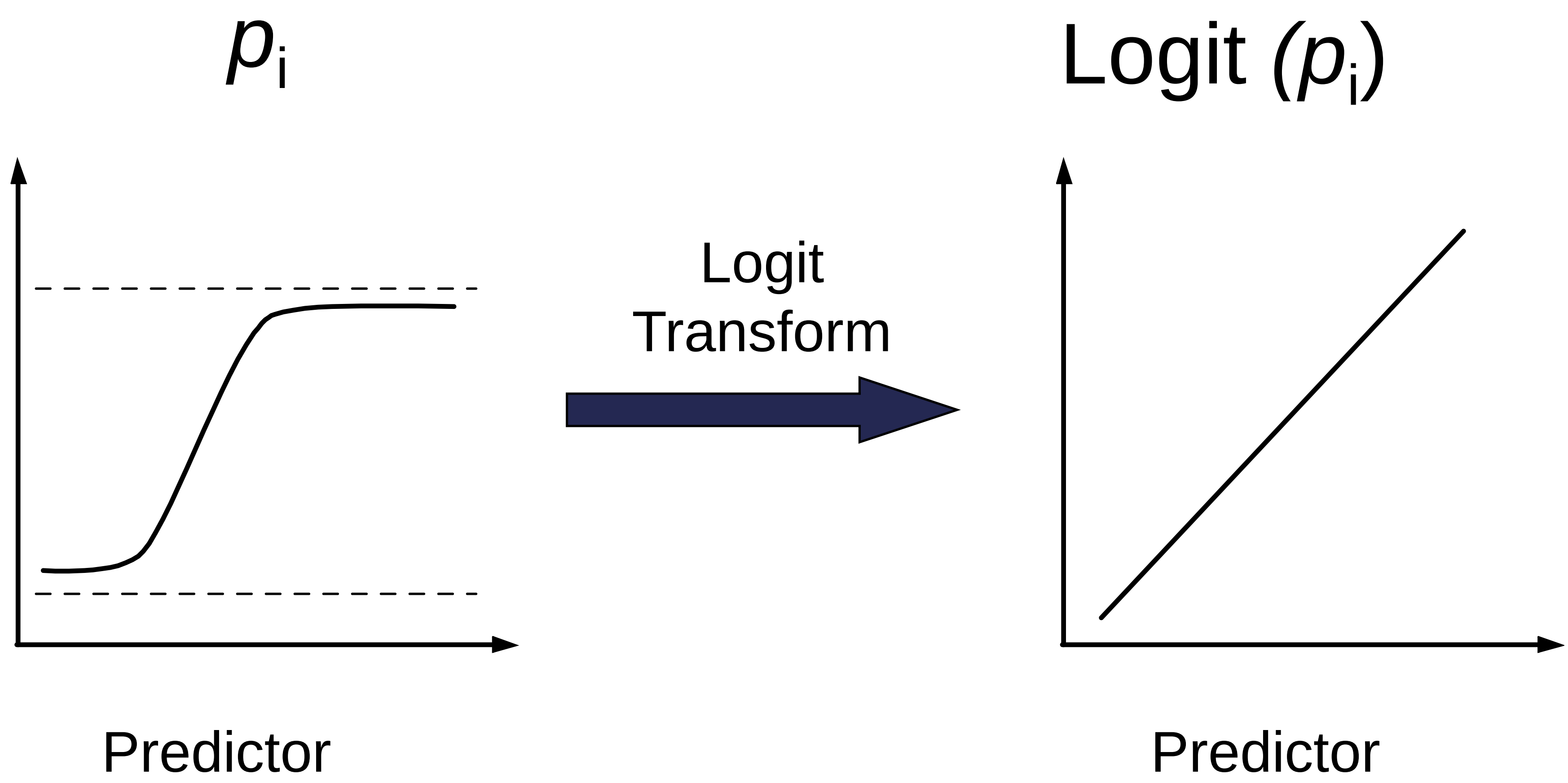


The Logit Link Transformation

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

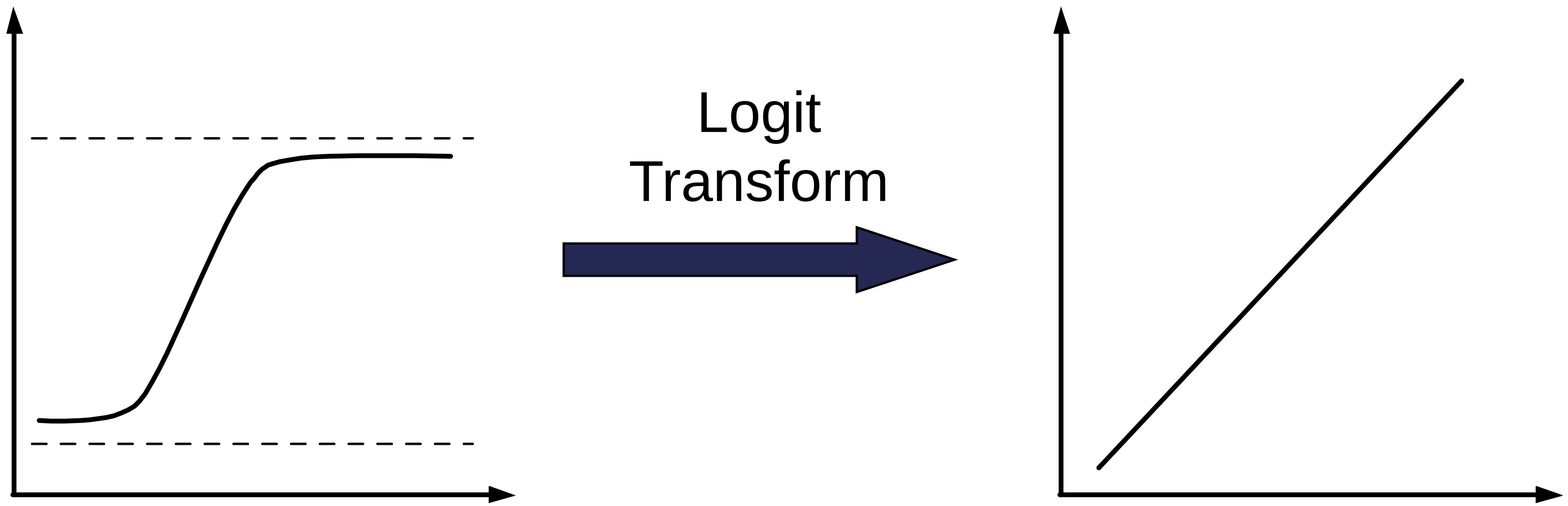
- To create a linear model, a link function (logit) is applied to the probabilities.
- The relationship between the parameters and the logits are linear.
- Logits unbounded.

The Logit Link Transformation



Assumption

p_i Logit (p_i)



Predictor Predictor

How do we evaluate?

Box-Tidwell Transformation

- Commonly used as a "test" for linearity of the X's relative to the logit in logistic regression models.
- Consider the following model:

$$\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i}^{\gamma_1} + \ldots + \beta_k X_{1i}^{\gamma_k}$$

- The Box-Tidwell transformation is a power transformation on the X's.
- Let's examine the case where $\gamma_i = 1$ for all i.

Box-Tidwell Transformation

$$\text{logit}(p_i) = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_k X_{1i}$$

$$\text{logit}(p_i) = \tilde{\beta}_0 + \tilde{\beta}_1 X_{1i} + \ldots + \tilde{\beta}_k X_{1i} + \hat{\delta}_1 X_{1i} \ln(X_{1i}) + \ldots + \hat{\delta}_k X_{ki} \ln(X_{ki})$$

$$\hat{\gamma}_i = 1 + \frac{\hat{\delta}_i}{\hat{\beta}_i}$$

Checking Assumptions – SAS

```
data lowbwt;
      set logistic.lowbwt;
      aloga = age*log(age);
      llogl = lwt*log(lwt);
run;

proc logistic data=lowbwt plots(only)=(oddsratio);
      class race(ref='white') / param=ref;
      model low(event='1') = age race lwt smoke aloga llogl /
                                 clodds=pl clparm=pl;
      title 'Modeling Low Birth Weight';
run;
quit;
```

Checking Assumptions – SAS

```
data lowbwt;
      set logistic.lowbwt;
      aloga = age*log(age);
      llogl = lwt*log(lwt);
run;

proc logistic data=lowbwt plots(only)=(oddsratio);
      class race(ref='white') / param=ref;
      model low(event='1') = age race lwt smoke aloga llogl /
                             clodds=pl clparm=pl;
      title 'Modeling Low Birth Weight';
run;
quit;
```

Checking Assumptions – SAS

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
age	1	1.0684	0.3013
race	2	7.7573	0.0207
lwt	1	0.0358	0.8499
smoke	1	7.3098	0.0069
aloga	1	1.1036	0.2935
llogl	1	0.0176	0.8945

Checking Assumptions – R

```
boxTidwell(low ~ age + lwt, data = bwt)
```

```
##      MLE of lambda Score Statistic (z) Pr(>|z|)
## age         3.9362             -0.7730   0.4395
## lwt        -4.3556              1.0178   0.3088
##
## iterations =  10
```

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!

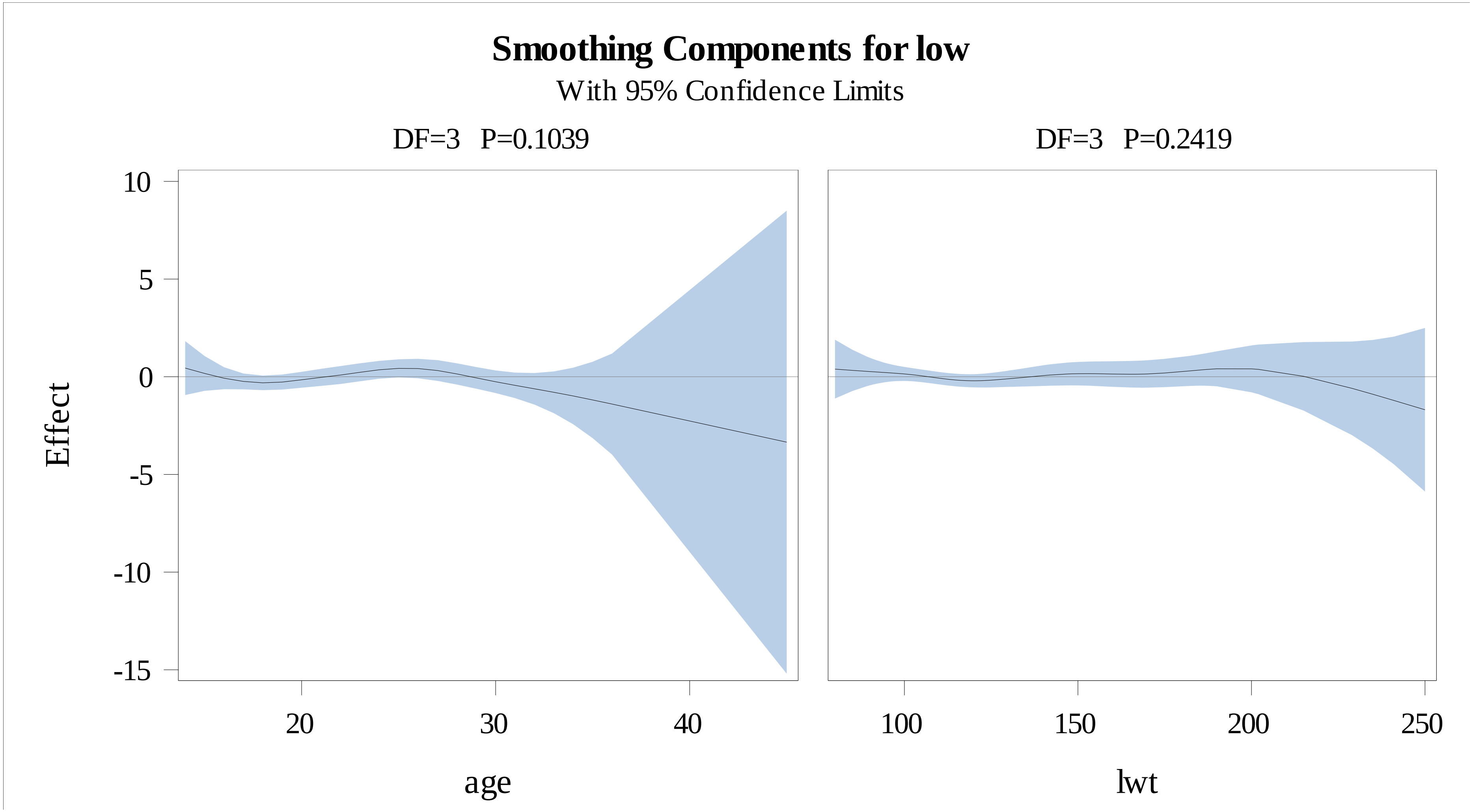
Checking Assumptions – SAS

```
proc gam data=logistic.lowbwt
         plots = components(clm commonaxes);
   class race(ref='white');
   model low(event='1') = spline(age, df=4)
                          spline(lwt, df=4)
                          param(smoke race)
                          / dist = binomial link = logit;
run;
```

Checking Assumptions – SAS

Smoothing Model Analysis Analysis of Deviance				
Source	**DF**	**Sum of Squares**	**Chi-Square**	**Pr > ChiSq**
Spline(age)	3.00000	6.162954	6.1630	0.1039
Spline(lwt)	3.00000	4.187660	4.1877	0.2419

Checking Assumptions – SAS



Checking Assumptions – R

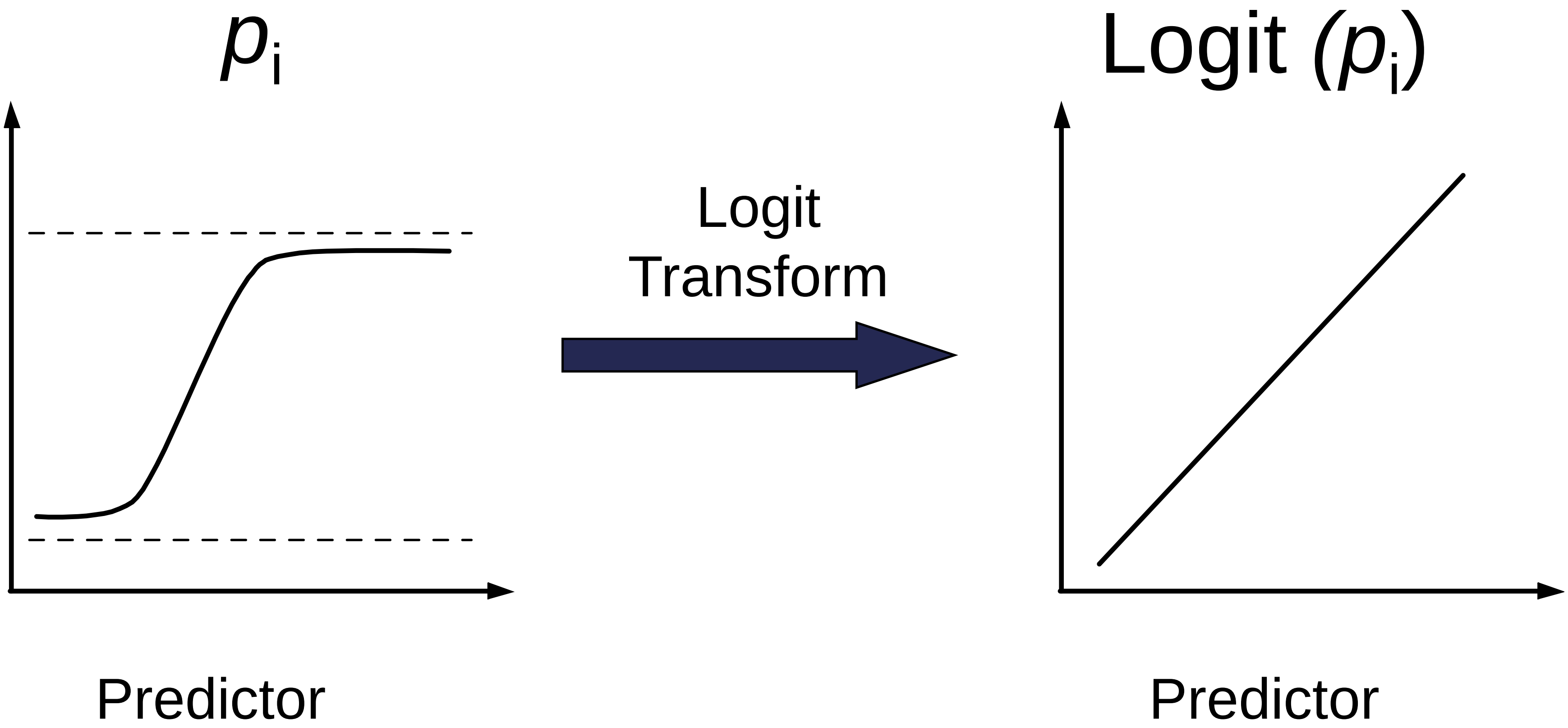
```
fit.gam <- gam(low ~ s(age) + s(lwt) + smoke + factor(race),
               data = bwt, family = binomial(link = 'logit'),
               method = 'REML')
summary(fit.gam)
```

Checking Assumptions – R

```
## Approximate significance of smooth terms:
##          edf Ref.df Chi.sq p-value
## s(age) 2.07   2.64  1.351  0.5494
## s(lwt) 1.00   1.00  3.764  0.0524 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0744   Deviance explained = 9.77%
## -REML = 109.46  Scale est. = 1         n = 189
```

COEFFICIENT INTERPRETATIONS

Unit Change in Predictor does…?



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

change in **Odds**

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

Odds Ratio from a Logistic Regression

- Estimated logistic regression model:

$$\text{logit}(p_i) = 0.332 + 1.054 * smoke + \cdots$$

- Estimated odds ratio (Smokers vs. Non-smokers):

$$\text{OR} = \frac{e^{0.332+1.054(1)+\cdots}}{e^{0.332+1.054(0)+\cdots}} = \frac{e^{0.332}e^{1.054}\ \ldots}{e^{0.332}\ \ldots} = e^{1.054} = 2.87$$

- Smokers have $\mathbf{100 * \left(e^{1.054} - 1\right)\% = 187\%}$ **higher expected odds** than non-smokers to have low birth weight babies.

Odds Ratio from a Logistic Regression

- Estimated logistic regression model:

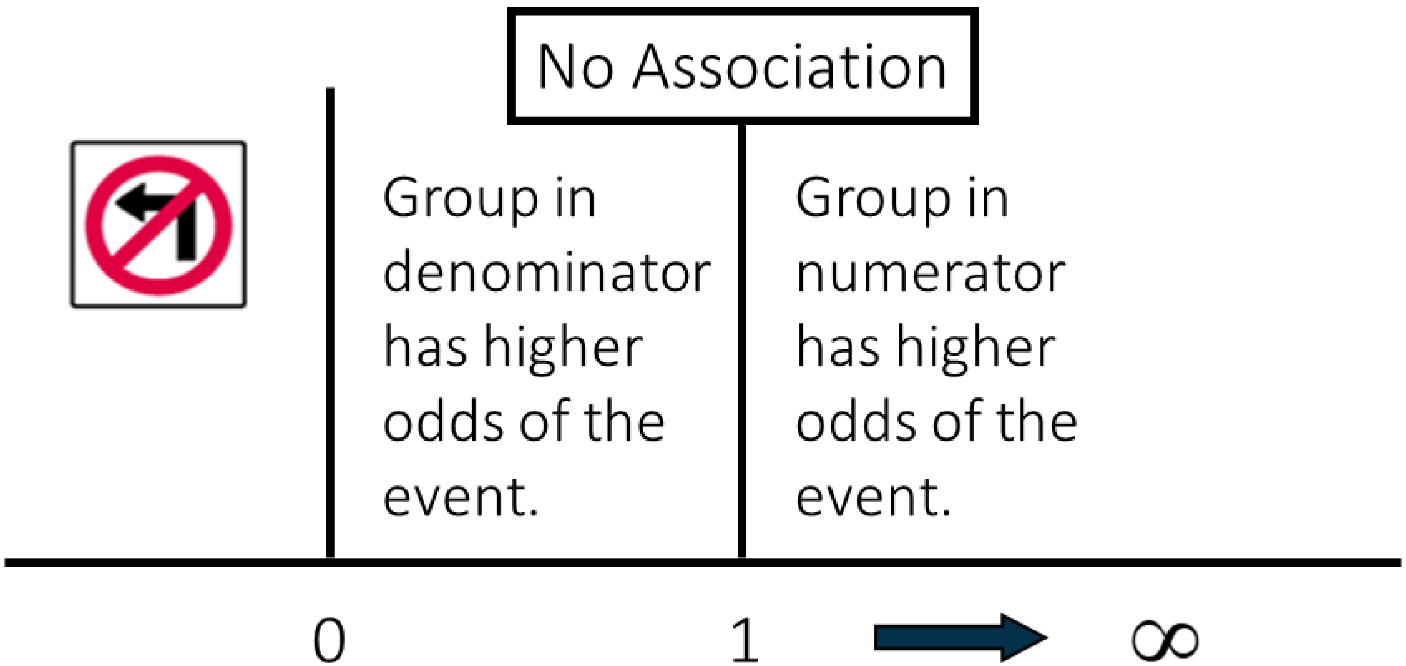
$$\text{logit}(p_i) = 0.332 - 0.022 * age + \cdots$$

- Estimated odds ratio (Additional Year of Age):

$$\text{OR} = \frac{e^{0.332-0.022(age+1)+\cdots}}{e^{0.332-0.022(age)+\cdots}} = e^{-0.022} = 0.98$$

- Every additional year of age **decreases the expected odds by 2%** to have low birth weight babies.

Properties of the Odds Ratio



Odds Ratios – SAS

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	**Unit**	**Estimate**	**95% Confidence Limits**	
age	1.0000	0.978	0.913	1.045
race black vs white	1.0000	3.427	1.247	9.629
race other vs white	1.0000	2.568	1.150	5.935
lwt	1.0000	0.988	0.974	0.999
smoke	1.0000	2.870	1.382	6.186

Odds Ratios – R

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

```
##                          2.5 %     97.5 %
## (Intercept)        4.7784821 0.5088761 50.9670681
## age                0.9777725 0.9131073  1.0445960
## lwt                0.9875525 0.9744679  0.9993613
## factor(smoke)1     2.8703634 1.3823204  6.1857015
## factor(race)other  0.7494552 0.2652201  2.1245479
## factor(race)white  0.2918045 0.1038416  0.8020311
```

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.

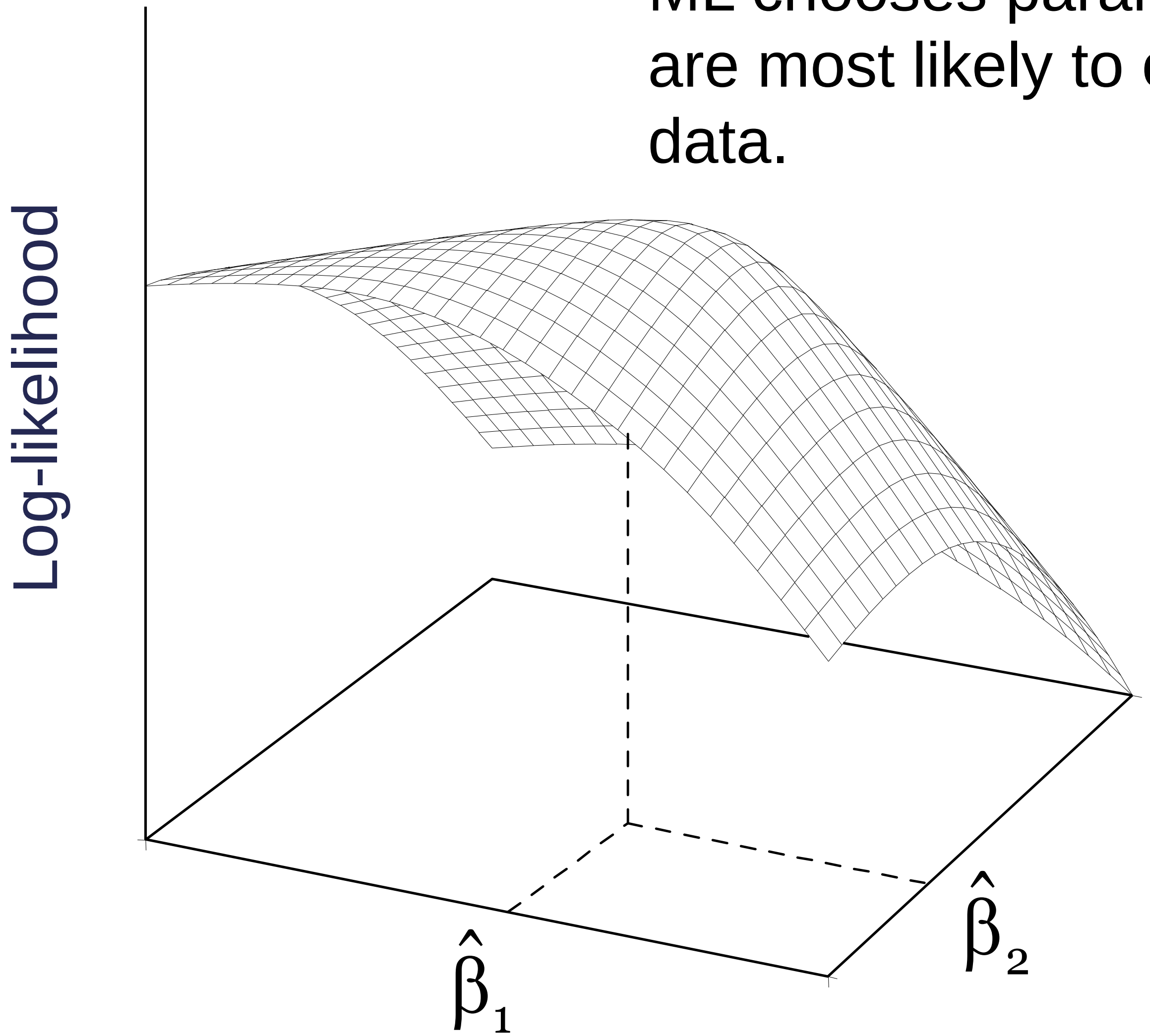
Maximum Likelihood Estimation

- 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!

Maximum Likelihood Estimation

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

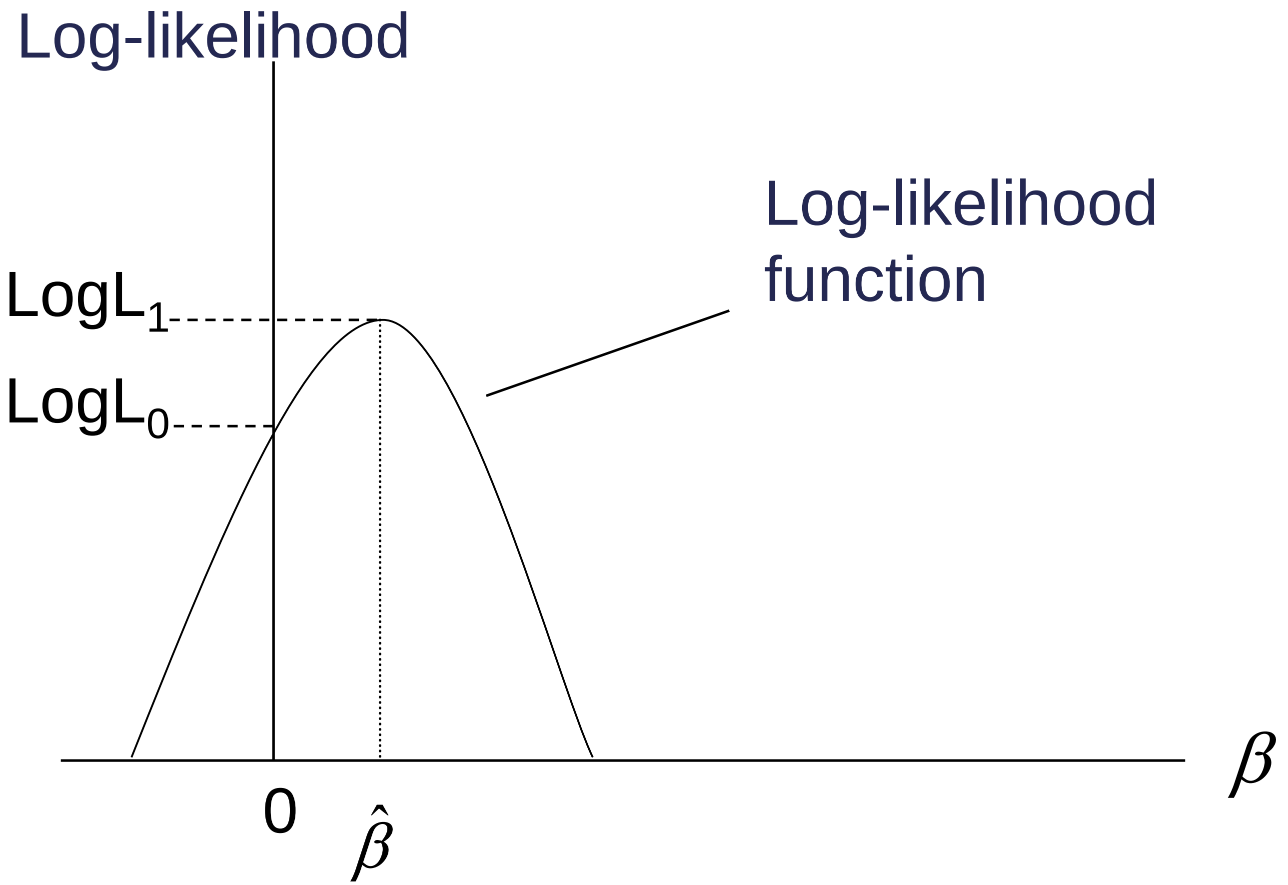


Likelihood Ratio Tests

- 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 – SAS

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	20.0948	5	0.0012
Score	18.6377	5	0.0022
Wald	16.4973	5	0.0056

Likelihood Ratio Test – R

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

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

```
## Analysis of Deviance Table
##
## Model 1: low ~ age + lwt + factor(smoke) + factor(race)
## Model 2: low ~ 1
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       183     214.58
## 2       188     234.67 -5  -20.095   0.0012 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
```

PREDICTED VALUES

Predicted Probabilities

p_i $\qquad\qquad$ Logit (p_i)

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

Predicted Values – SAS

```
proc logistic data=logistic.lowbwt plots(only)=(oddsratio);
   class race(ref='white') smoke(ref='0') / param=ref;
   model low(event='1') = age race lwt smoke /
                          clodds=pl clparm=pl;
   title 'Modeling Low Birth Weight';
   score data=newbw out=bw_scored;
run;
quit;
```

Predicted Values – SAS

Obs	Race	Low	Age	Lwt	Smoke	F_low	I_low	P_0	P_1
1	white	1	21	110	0	1	0	0.8202	0.1798
2	black	0	40	120	0	0	0	0.6981	0.3109
3	other	1	31	130	1	1	1	0.4988	0.5012
4	white	0	28	140	1	0	0	0.7303	0.2697
5	black	1	35	100	0	1	0	0.6166	0.3834

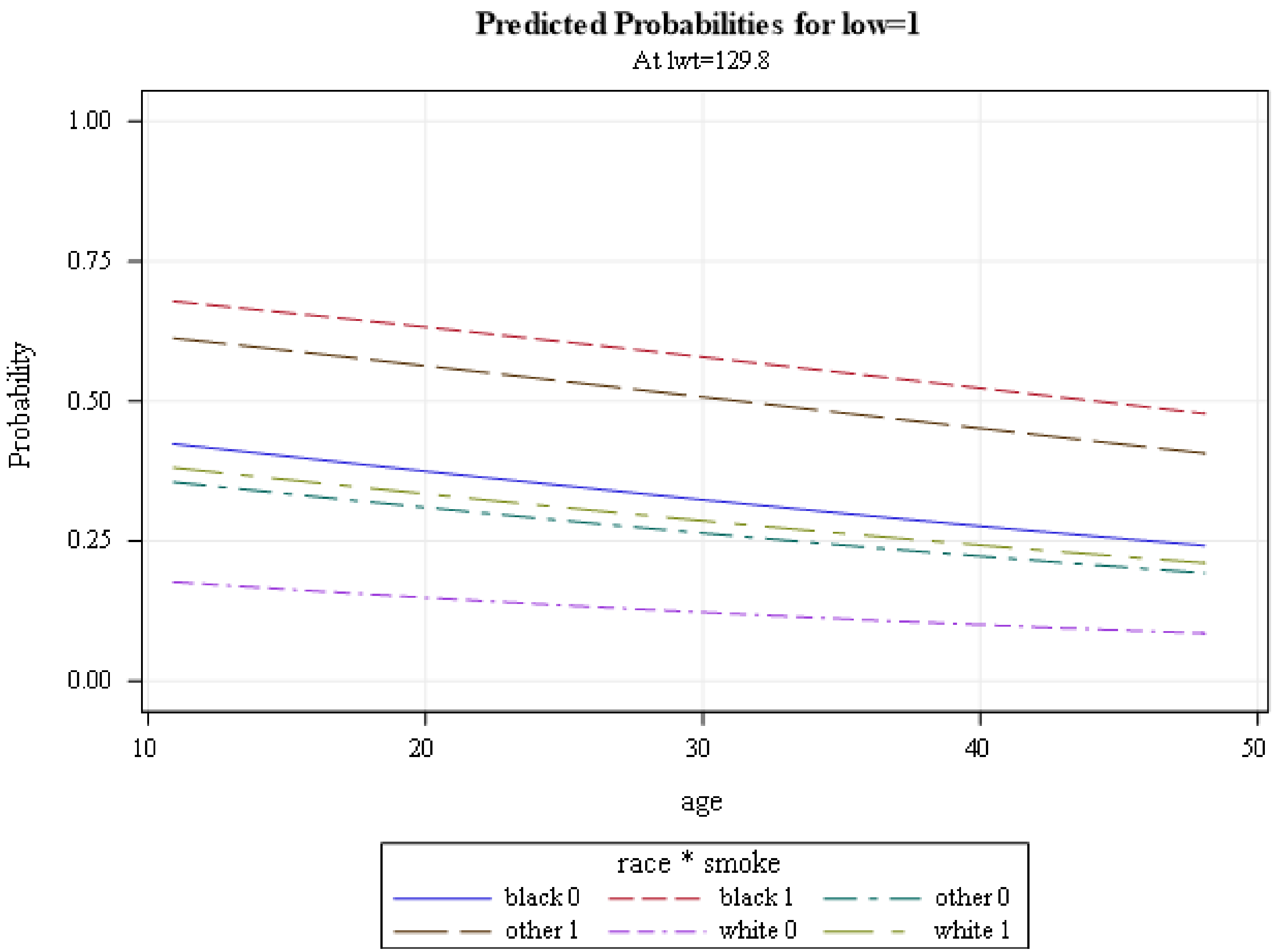
Predicted Values – SAS

Obs	Race	Low	Age	Lwt	Smoke	F_low	I_low	P_0	P_1
1	white	1	21	110	0	1	0	0.8202	0.1798
2	black	0	40	120	0	0	0	0.6981	0.3109
3	other	1	31	130	1	1	1	0.4988	0.5012
4	white	0	28	140	1	0	0	0.7303	0.2697
5	black	1	35	100	0	1	0	0.6166	0.3834

Predicted Values – SAS

Obs	Race	Low	Age	Lwt	Smoke	F_low	I_low	P_0	P_1
1	white	1	21	110	0	1	0	0.8202	0.1798
2	black	0	40	120	0	0	0	0.6981	0.3109
3	other	1	31	130	1	1	1	0.4988	0.5012
4	white	0	28	140	1	0	0	0.7303	0.2697
5	black	1	35	100	0	1	0	0.6166	0.3834

Predicted Probability Plot – SAS



Predicted Values – R

```
predict(logit.model, newdata = newbw, type = "response")
```

```
##         1         2         3         4         5
## 0.1798424 0.3019376 0.5012475 0.2697100 0.3833902
```

Predicted Probability Plot – R

```
visreg(logit.model, "lwt", by = "race", scale = "response",
       cond = list(smoke = 0, lwt = 130),
       overlay = TRUE,
       xlab = 'Mother`s Weight',
       ylab = 'Low Birth Weight')
```

Predicted Probability Plot – R