REVIEW OF LOGISTIC REGRESSION

Dr. Aric LaBarr

Institute for Advanced Analytics

MATH REVIEW

Odds vs. Probability

- **Odds** is the ratio of events to non-events:

$$Odds = \frac{\#yes}{\#no}$$

- **Probability** is the ratio of event to the total number of outcomes:

$$p = \frac{\#yes}{\#yes + \ \#no}$$

- **Odds** and **Probability** are related:

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

Probability versus Odds of an Outcome

	No Buy	Buy	Total
No Checking	30	54	**84**
Checking	291	125	**416**
Total	**321**	**179**	**500**

Probability versus Odds of an Outcome

	No Buy	Buy	Total
No Checking	30	54	**84**
Checking	291	125	**416**
Total	**321**	**179**	**500**

Probability of **NO BUY** in **Checking** account customers $= \frac{291}{416} = 0.70$

Probability versus Odds of an Outcome

	No Buy	Buy	Total
No Checking	30	54	**84**
Checking	291	125	**416**
Total	**321**	**179**	**500**

Probability of **BUY** in **Checking** account customers $= \frac{125}{416} = 0.30$

Probability versus Odds of an Outcome

	No Buy	Buy	Total
No Checking	30	54	**84**
Checking	291	125	**416**
Total	**321**	**179**	**500**

Odds of **BUY** in **Checking** account customers

$$= \frac{\text{Prob. of Buy}}{\text{Prob. of No Buy}} = \frac{0.30}{0.70} = 0.43$$

Odds Ratio

- **Odds Ratio** indicates how likely (in terms of odds) an event is for one group relative to another:

$$OR = \frac{Odds_A}{Odds_B}$$

- Since odds are always non-negative, so are odds ratios
  - OR > 1 → Event **more likely for A than for B**
  - OR < 1 → Event **more likely for B than for A**
  - OR = 1 → Event **equally likely in each group**

Probability versus Odds of an Outcome

	No Buy	Buy	Total
No Checking	30	54	84
Checking	291	125	416
Total	321	179	500

Odds of **BUY** in **No Checking** $= 1.77$

Odds of **BUY** in **Checking** $= 0.43$

Odds Ratio: No Checking to Checking $= \frac{1.77}{0.43} = 4.12$

Odds Ratio

Odds of **BUY** in **No Checking** $= 1.77$

Odds of **BUY** in **Checking** $= 0.43$

Odds Ratio: No Checking to Checking $= \dfrac{1.77}{0.43} = 4.12$

Non-Checking account customers have **4.12 times the odds** of buying the insurance product as compared to checking account customers.

Relative Risk

- **Relative Risk** indicates how likely (in terms of probability) an event is for one group relative to another:

$$RR = \frac{p_A}{p_B}$$

- Since probabilites are always non-negative, so are relative risks
  - RR > 1 → Event **more likely for A than for B**
  - RR < 1 → Event **more likely for B than for A**
  - RR = 1 → Event **equally likely in each group**

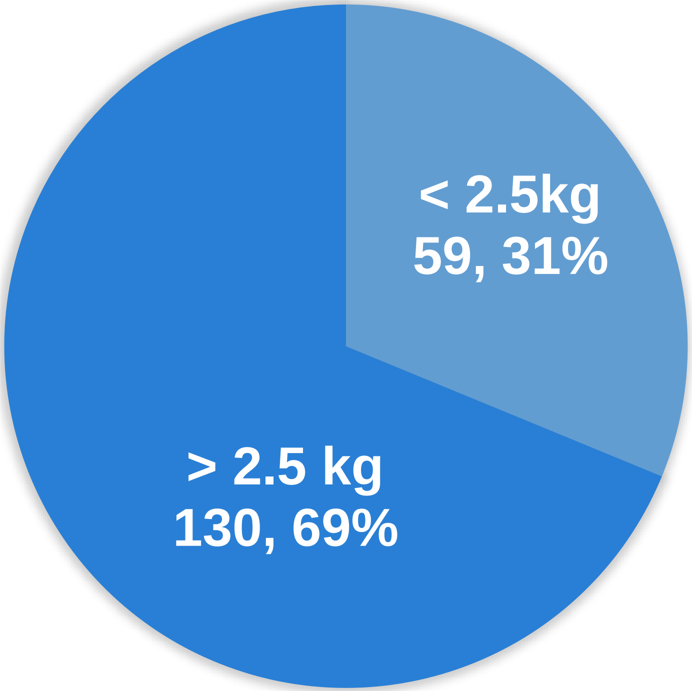
Math for Logistic Regression

- The following are rules involving the exponential function and natural logarithm:
  - $e^a > 0$ for any number *a*
  - $e^{a+b} = e^a e^b$, and $e^{a-b} = \frac{e^a}{e^b}$
  - $\log(a)$ can be any number, but $a > 0$
    - $\log(a) = -\infty$ if $a = 0$
    - $\log(a)$ **does not exist** if $a < 0$
  - $\log(a \times b) = \log(a) + \log(b)$, and $\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$
  - $\log(e^a) = a$, and $e^{\log(a)} = a$
  - $a^{-1} = \frac{1}{a}$

BINARY LOGISTIC REGRESSION REVIEW

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

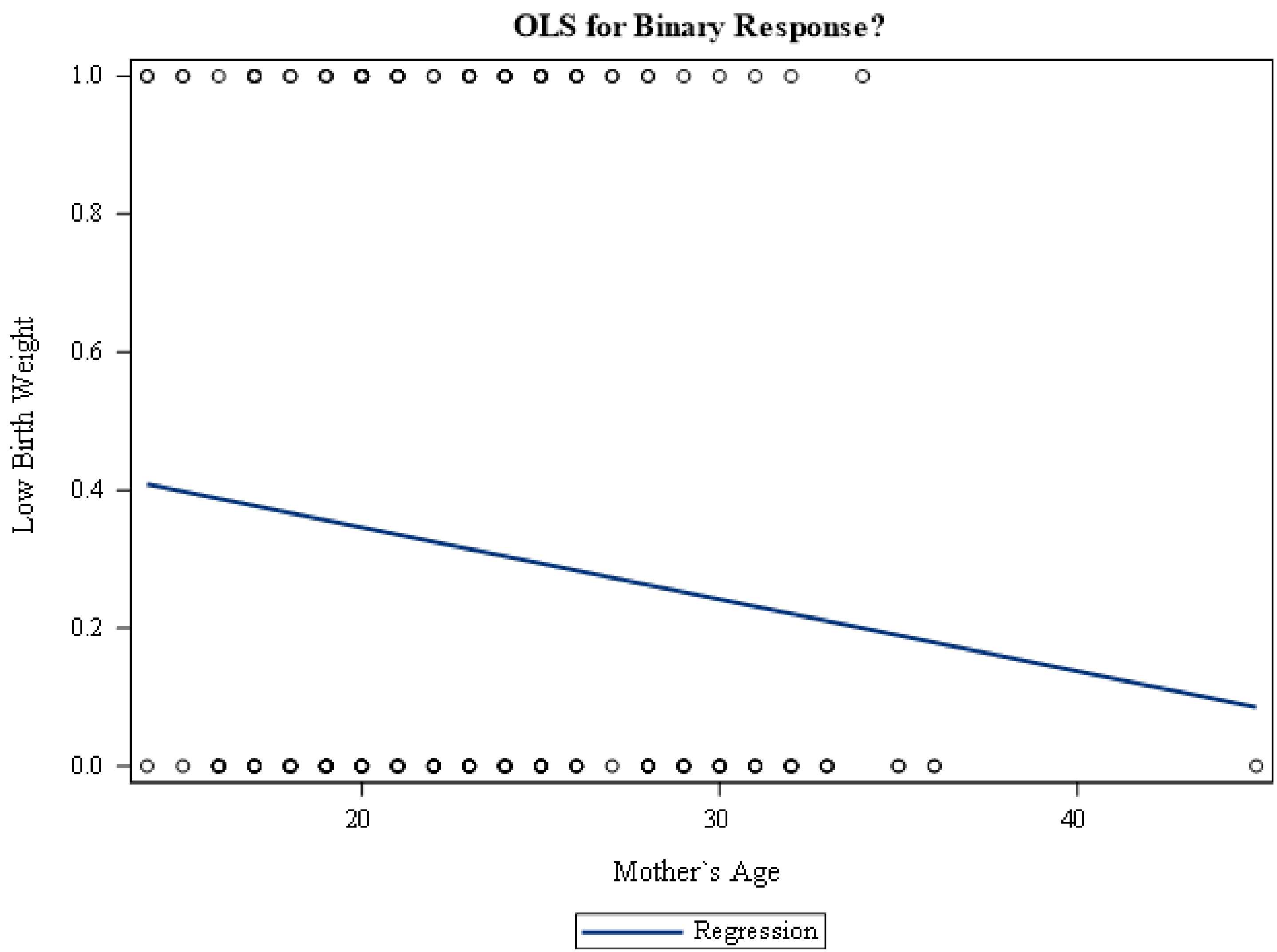
- Model the association between various factors and child being born with low birth weight (< 2.5kg)
- Predictors:
  - **age:** mother's age (years)
  - **lwt:** mother's weight at last menstrual period (lbs)
  - **smoke:** mother's smoking status during pregnancy
  - **race:** mother's race (1=White,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

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?

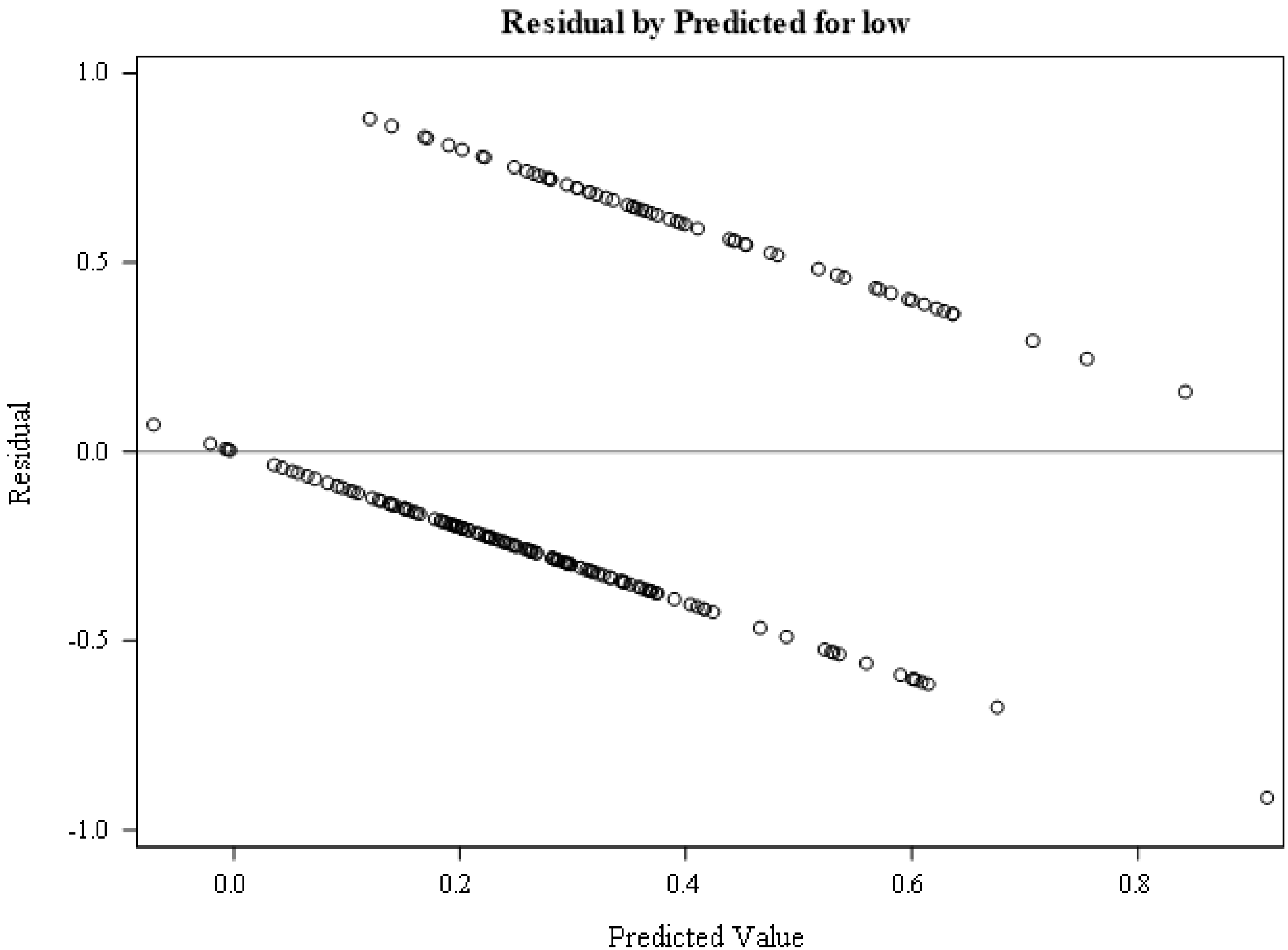
OLS Regression Plot



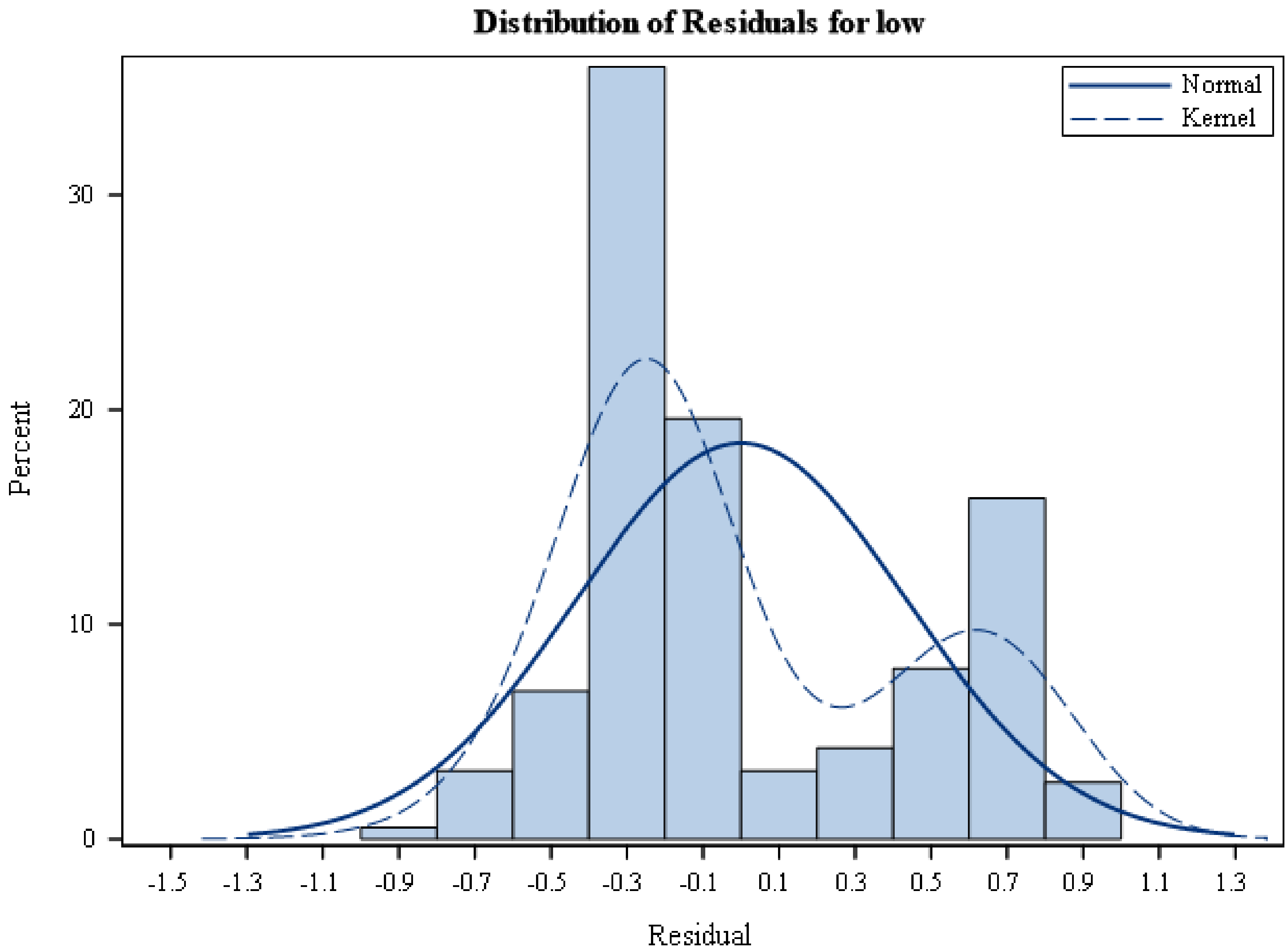
Assumptions for OLS Regression

- 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.

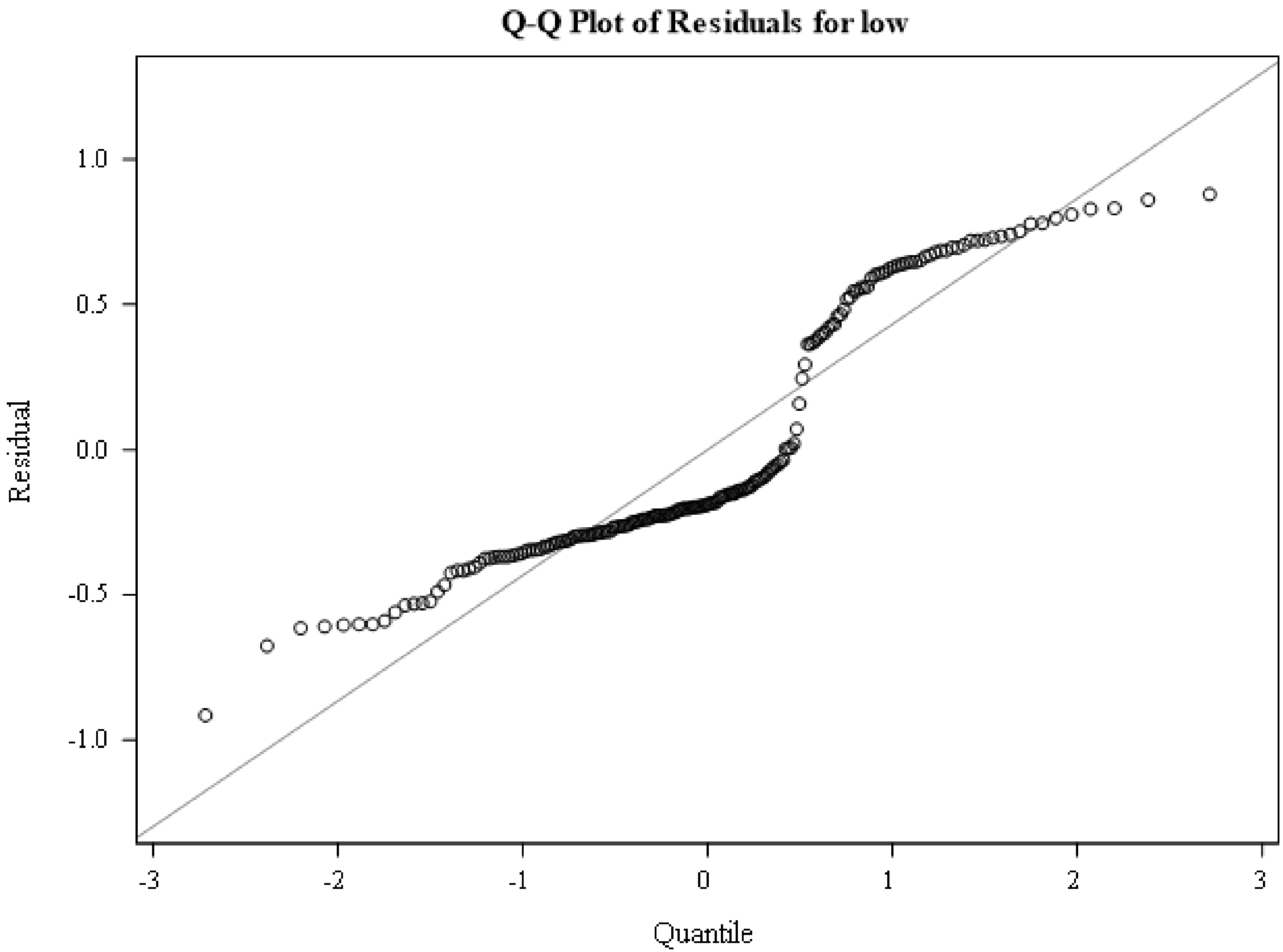
Assumptions for OLS Regression



Assumptions for OLS Regression



Assumptions for OLS Regression



Linear Probability Model

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

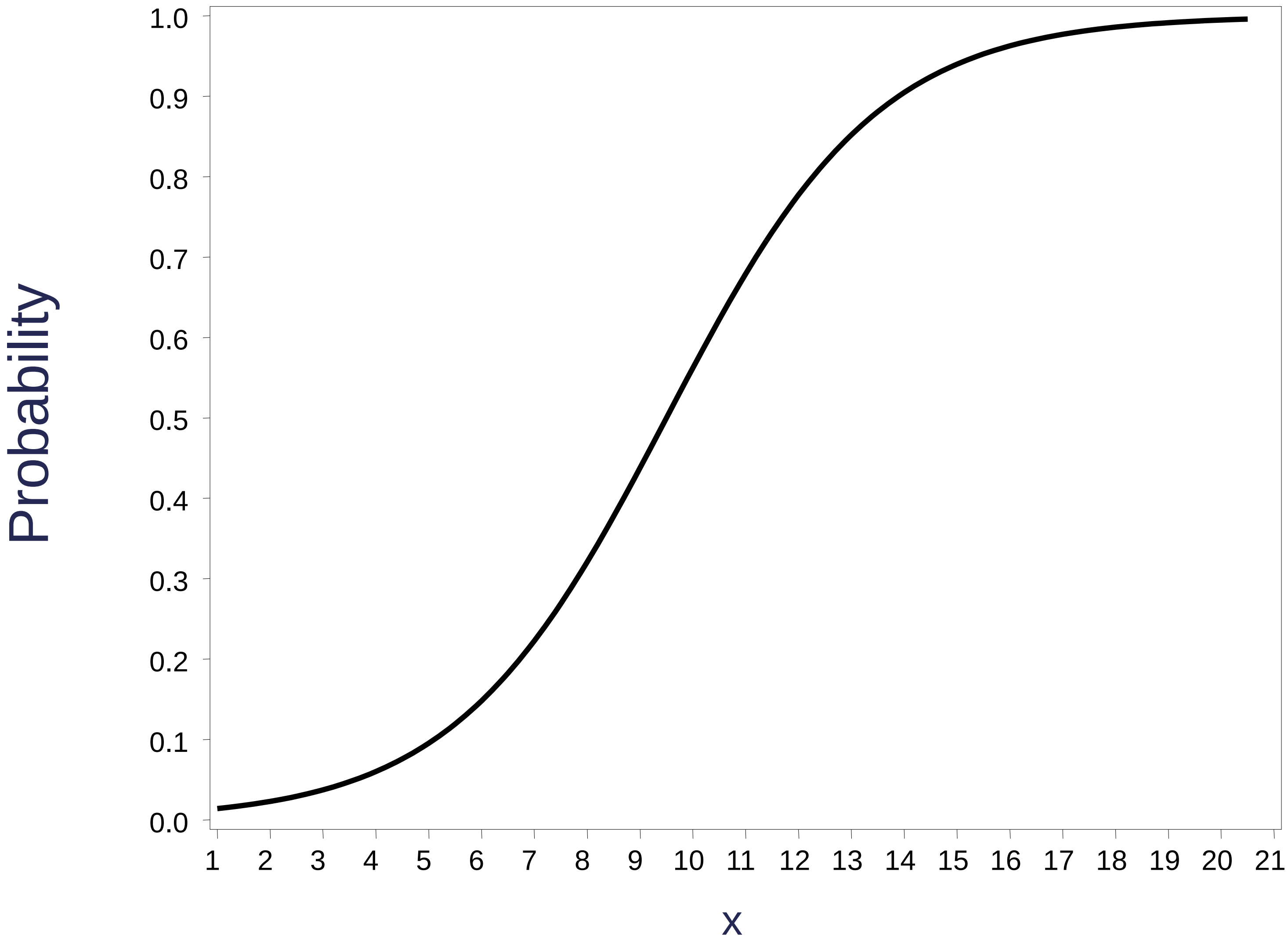
- Problems:
  - Probabilities are bounded, but linear functions can take on any value. (How do you interpret a predicted value of -0.4 or 1.1?)
  - The relationship between probabilities and X is usually nonlinear. Example, one unit change in X will have different effects when the probability is near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

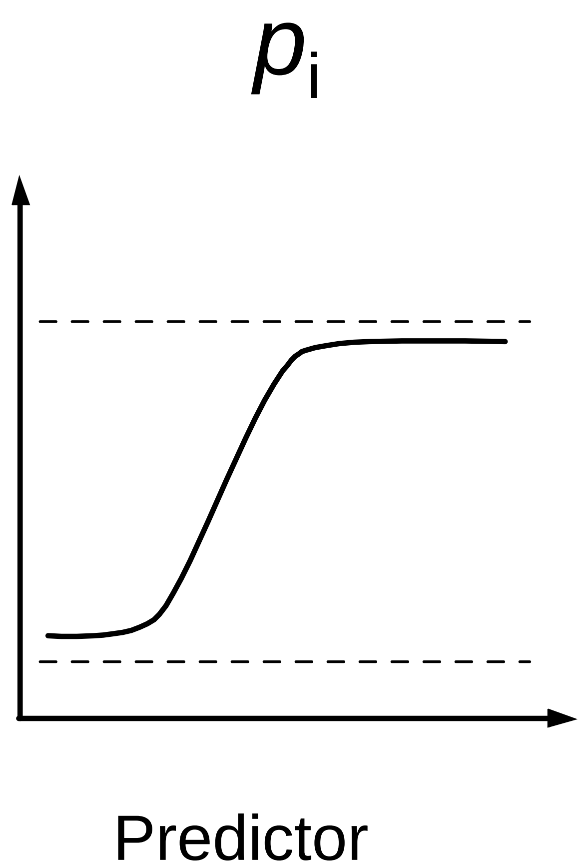


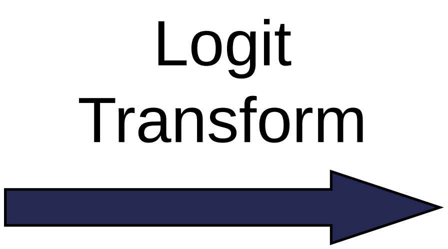
The Logit Link Transformation

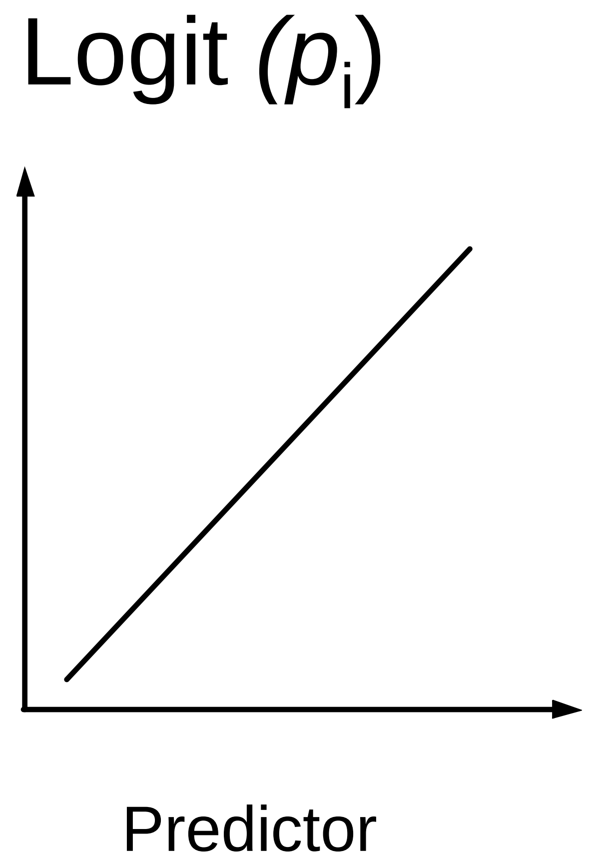
$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots \beta_k x_{k,i}$$

- 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







CATEGORICAL INPUTS

Reference Coding

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Reference coding** is a common way to code categorical variables.
- 2 Category Example (A, B):

$$x = \begin{cases} 1 & \text{if A} \\ 0 & \text{if B} \end{cases}$$

- 3 Category Example (A, B, C):

	x_1	x_2
A	1	0
B	0	1
C	0	0

Reference Coding

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Reference** coding is a common way to code categorical variables.
- 3 Category Example (A, B, C):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Average difference between category A and C.

	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$
A	1	0
B	0	1
C	0	0

Reference Coding

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Reference coding** is a common way to code categorical variables.
- 3 Category Example (A, B, C):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Average difference between category B and C.

	x_1	x_2
A	1	0
B	0	1
C	0	0

Effects Coding

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Effects coding** is another common way to code categorical variables.
- 2 Category Example (A, B):

$$x = \begin{cases} 1 & \text{if A} \\ \mathbf{-1} & \text{if B} \end{cases}$$

- 3 Category Example (A, B, C):

	x_1	x_2
A	1	0
B	0	1
C	**-1**	**-1**

Effects Coding

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Effects coding** is another common way to code categorical variables.
- 3 Category Example (A, B, C):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Average difference between category A and the overall average of categories **A, B, & C**.

	x_1	x_2
A	1	0
B	0	1
C	**-1**	**-1**

Effects Coding

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Effects coding** is another common way to code categorical variables.
- 3 Category Example (A, B, C):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Average difference between category B and the overall average of categories **A, B, & C**.

	x_1	x_2
A	1	0
B	0	1
C	**-1**	**-1**

BINARY LOGISTIC REGRESSION IN SAS

Logistic Regression Model – SAS

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

Logistic Regression Model – SAS

Modeling Low Birth Weight
The LOGISTIC Procedure

Model Information	
Data Set	LOGISTIC.LOWBWT
Response Variable	low
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	189
Number of Observations Used	189

Logistic Regression Model – SAS

Response Profile		
Ordered Value	**low**	**Total Frequency**
1	0	130
2	1	59

Probability modeled is low='1'.

Class Level Information			
Class	**Value**	**Design Variables**	
race	**black**	1	0
	other	0	1
	white	0	0

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

Logistic Regression Model – SAS

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	236.672	226.577
SC	239.914	246.028
-2 Log L	234.672	214.577

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

Logistic Regression Model – SAS

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
age	1	0.4326	0.5107
race	2	7.8419	0.0198
lwt	1	3.8470	0.0498
smoke	1	7.6991	0.0055

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	0.3324	1.1077	0.0900	0.7641
age		1	-0.0225	0.0342	0.4326	0.5107
race	**black**	1	1.2316	0.5171	5.6718	0.0172
race	**other**	1	0.9432	0.4162	5.1351	0.0234
lwt		1	-0.0125	0.00639	3.8470	0.0498
smoke		1	1.0544	0.3800	7.6991	0.0055

Logistic Regression Model – SAS

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
age	1	0.4326	0.5107
race	2	7.8419	0.0198
lwt	1	3.8470	0.0498
smoke	1	7.6991	0.0055

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	0.3324	1.1077	0.0900	0.7641
age		1	-0.0225	0.0342	0.4326	0.5107
race	**black**	1	1.2316	0.5171	5.6718	0.0172
race	**other**	1	0.9432	0.4162	5.1351	0.0234
lwt		1	-0.0125	0.00639	3.8470	0.0498
smoke		1	1.0544	0.3800	7.6991	0.0055

Logistic Regression Model – SAS

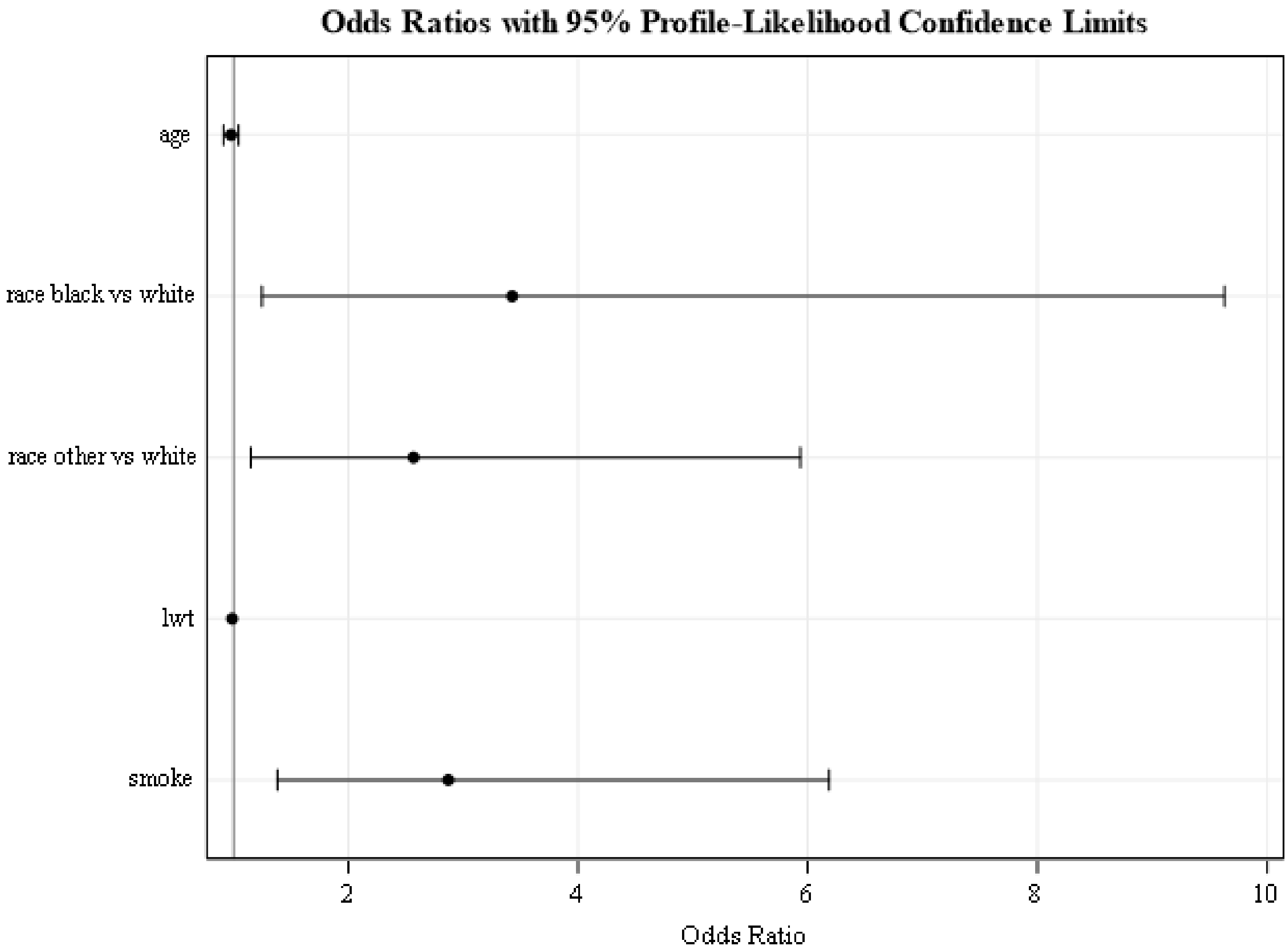
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	68.4	**Somers' D**	0.367
Percent Discordant	31.6	**Gamma**	0.368
Percent Tied	0.0	**Tau-a**	0.159
Pairs	7670	**c**	0.684

Parameter Estimates and Profile-Likelihood Confidence Intervals				
Parameter		**Estimate**	**95% Confidence Limits**	
Intercept		0.3324	-1.8092	2.5609
age		-0.0225	-0.0909	0.0436
race	**black**	1.2316	0.2206	2.2648
race	**other**	0.9432	0.1401	1.7809
lwt		-0.0125	-0.0259	-0.00064
smoke		1.0544	0.3238	1.8222

Logistic Regression Model – 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

Logistic Regression Model – SAS



Logistic Regression Model – SAS