

# Multiple Logistic Regression for Dichotomous Responses Edpsy/Psych/Soc 589

Carolyn J. Anderson

Department of Educational Psychology



ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

# Outline

## Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

In last set of notes:

- Review and Some Uses & Examples.
- Interpreting logistic regression models.
- Inference for logistic regression.
- Model checking.

---

This set of notes covers:

- Logit models for qualitative explanatory variables.
- Multiple logistic regression.
- The Tale of the Titanic
- Sample size & power.

# Qualitative Explanatory Variables

Explanatory variables can be

- Continuous
- Discrete – nominal
- Discrete – ordinal
- Continuous and Discrete (or “mixed”).

We'll now consider the case of discrete variables and mixed when we discuss multiple logistic regression.

For example, in the high school and beyond data set we could look at whether students who attend academic versus non-academic programs differed in terms of

- School type (public or private).
- Race (4 levels).
- Career choice (11 levels).
- SES level (3 levels).

Overview

Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic Regression

# Yes, HSB data yet again

For purposes of illustration, we'll use the following data:

SES Level	School Type	Program Type		$n_i$
		Non-Academic	Academic	
Low	public	91	40	131
	private	4	4	8
Middle	public	138	111	249
	private	14	36	50
High	public	44	82	126
	private	1	35	36
				600

We can incorporate nominal discrete variables by creating **Dummy Variables** and include them in our model.

Overview

Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic Regression

# Dummy Variables

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

For **School type**

$$\begin{aligned} X_1 &= 1 && \text{if public} \\ &= 0 && \text{if private} \end{aligned}$$

For **SES:**

$$\begin{aligned} S_1 &= 1 && \text{if low} \\ &= 0 && \text{otherwise} \\ S_2 &= 1 && \text{if middle} \\ &= 0 && \text{otherwise} \end{aligned}$$

Our **logit model** is

$$\text{logit}(\pi) = \alpha + \beta_1 x_1 + \beta_2 s_1 + \beta_3 s_2$$

# HSB model: $\text{logit}(\pi) = \alpha + \beta_1 x_1 + \beta_2 s_1 + \beta_3 s_2$

This model has “main” effects for school type (i.e.,  $\beta_1$ ) and SES (i.e.,  $\beta_2$  and  $\beta_3$ ).

With our dummy variables defined as

$x_1 = 1$  for public and  $= 0$  for private.

$s_1 = 1$  for low SES and  $= 0$  for middle or high.

$s_2 = 1$  for middle SES and  $= 0$  for low and high.

For each combination of the explanatory variables,

SES Level	School Type	$x_1$	$s_1$	$s_2$	$\text{logit}(\pi) = \log(\text{academic/nonacademic})$
Low	public	1	1	0	$\alpha + \beta_1 + \beta_2$
	private	0	1	0	$\alpha + \beta_2$
Middle	public	1	0	1	$\alpha + \beta_1 + \beta_3$
	private	0	0	1	$\alpha + \beta_3$
High	public	1	0	0	$\alpha + \beta_1$
	private	0	0	0	$\alpha$

What do the parameters  $\beta_1$ ,  $\beta_2$  and  $\beta_3$  mean?

# Interpreting $\beta$ 's

$$\text{logit}(\pi) = \alpha + \beta_1 x_1 + \beta_2 s_1 + \beta_3 s_2$$

$\exp(\beta_1)$  = the conditional odds ratio between program type and given SES.

For example, for low SES

$$\begin{aligned} \frac{(\text{odds academic})|_{\text{public, low}}}{(\text{odds academic})|_{\text{private, low}}} &= \frac{\exp(\alpha + \beta_1 + \beta_2)}{\exp(\alpha + \beta_2)} \\ &= \frac{e^\alpha e^{\beta_1} e^{\beta_2}}{e^\alpha e^{\beta_2}} \\ &= e^{\beta_1} \end{aligned}$$

Since this does not depend on SES level (i.e.,  $\beta_2$  or  $\beta_3$ ),

$$\exp(\beta_1) = \frac{(\text{odds academic})|_{\text{public}}}{(\text{odds academic})|_{\text{private}}} (\text{SES})$$

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

# Interpreting the Other $\beta$ 's

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

- $\exp(\beta_2)$  = the conditional odds ratio between program type and low versus high SES given school type,

$$\exp(\beta_2) = e^{\beta_2} = \frac{(\text{odds academic})|_{\text{low}}}{(\text{odds academic})|_{\text{high}}} (\text{school type})$$

- $\exp(\beta_3)$  = the conditional odds ratio between program type and middle versus high SES given school type,

$$\exp(\beta_3) = e^{\beta_3} = \frac{(\text{odds academic})|_{\text{middle}}}{(\text{odds academic})|_{\text{high}}} (\text{school type})$$

- $\exp(\beta_2 - \beta_3)$  = the conditional odds ratio between program type and low versus middle SES given school type, or

$$\exp(\beta_2) / \exp(\beta_3) = e^{\beta_2 - \beta_3} = \frac{(\text{odds academic})|_{\text{low}}}{(\text{odds academic})|_{\text{middle}}} (\text{school type})$$

# Patterns of Association in 3-way Tables

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

- **Question:** What can we say about the association in a 3–way table when the conditional odds ratios do not depend on the level of the third variable?
- **Answer: Homogeneous Association** — So if a logit model with only main effects for the (qualitative) explanatory variables fits a 3–way table, then we know that the table displays homogeneous association.

Therefore, we can use estimated parameters to compute estimates of common odds ratios.

- **Question:** What would the model look like if the program type and SES were conditionally independent given school type?
- **Answer: Conditional independence** means that the conditional odds ratios of program type and SES for each level of school type are equal; that is,

$$\beta_2 = \beta_3 = 0$$

So the logit model is:  $\text{logit}(\pi) = \alpha + \beta_1 x_1.$

# Results

Using *SAS/GENMOD* or *LOGISTIC* we get the following:

Statistic	<i>df</i>	Value	<i>p</i> -value
$X^2$	2	3.748	.10
$G^2$ (deviance)	2	4.6219	.15
Log Likelihood		-375.3239	

The model looks like it fits OK; that is, the data display homogeneous association.

The estimated parameters, ASE and Wald statistics:

Variable	Estimate	<i>ASE</i>	Wald	<i>p</i> -value	Odds ratio $\exp(\beta)$
Intercept	$\hat{\alpha} = 2.1107$	.3060	47.5665	< .001	
School type ( $x_1$ )	$\hat{\beta}_1 = -1.3856$	.2792	24.6228	< .001	.25
Low SES ( $s_1$ )	$\hat{\beta}_2 = -1.5844$	.2578	37.7751	< .001	.21
Middle SES ( $s_2$ )	$\hat{\beta}_3 = -.9731$	.2152	20.4544	< .001	.38

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

# What the Results Mean

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

The estimated model:

$$\text{logit}(\hat{\pi}_i) = 2.1107 - 1.3856x_{1i} - 1.5844s_{1i} - .9731s_{2i}$$

Questions:

- Are Program type and school type conditionally independent given SES?
- Are Program type and SES conditionally independent given school type?

# Tests for Patterns of Association

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

- Breslow-Day statistic = 3.872,  $df = 2$ ,  $p = .144$
- CMH statistic for conditional independence of program type and school type given SES equals

$$\text{CMH} = 27.008, \quad df = 1, \quad p < .001$$

- The conditional likelihood ratio test of effect for  $\beta_1$   
( $H_0 : \beta_1 = 0$ )

$$G^2 = 14.37, \quad df = 1, \quad p < .001$$

- Testing conditional independence of program type and SES using a conditional likelihood ratio test (i.e.,  
 $H_0 : \beta_2 = \beta_3 (= 0)$ )

$$G^2 = 21.14, \quad df = 2, \quad p < .001$$

- The Mantel-Haentzel estimate of the common odds ratio between program type and school type given SES is

$$.238 \quad \text{or} \quad 1/.238 = 4.193$$

- and the one based on the logit model is

$$\exp(\hat{\beta}_1) = \exp(-1.3856) = .250 \quad \text{or} \quad 1/.250 = 4.00$$

# ANOVA–Type Representation

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

- When an explanatory variable has only 2 levels (e.g., school type), we only need a single dummy variable.
- When an explanatory variables has more than 3 levels, say  $I$  levels, then we need  $I - 1$  dummy variables (e.g., for SES we needed  $3-1=2$  dummy variables).
- When explanatory variables are discrete
  - ◆ We often call them “factors”.
  - ◆ Rather than explicitly writing out all the dummy variables, we represent the model as

$$\text{logit}(\pi) = \alpha + \beta_i^X + \beta_k^Z$$

where

- $\beta_i^X$  is the parameter for the  $i$ th level of variable  $X$ .
- $\beta_k^Z$  is the parameter for the  $k$ th level of variable  $Z$ .
- Conditional independence of (say)  $Y$  and  $X$  given  $Z$  would mean that  $\beta_1^X = \beta_2^X = \dots = \beta_I^X$ .
- There is a redundancy in the parameters; that is, if  $X$  has  $I$  levels, then you only need  $I - 1$  parameters.

# Identification Constraints

Are needed to estimate the parameters of the model.

The constraints do not effect

- Estimated fitted/predicted values of  $\pi$  (or  $\text{logit}(\pi)$ ), and therefore do not effect goodness of fit statistics or residuals.
- Estimated odds ratios.

The constraints do effect that actual values of the parameter estimates.

Typical constraints are

- Fix one value of a set, e.g.,  $\beta_1 = 0$  or  $\beta_I = 0$  (SAS/GENMOD)
- Fix sum equal to constant, usually 0. e.g.,  $\sum_{i=1}^I \beta_i = 0$ .

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

# Identification Constraints: example

Overview

Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic Regression

Term	Dummy code		Effect code
	Fix First	Fix Last	Zero Sum
	$\beta_1 = 0$	$\beta_I = 0$	$\sum_i \beta_i = 0$
Intercept	-.8593	2.1107	.5654
Public	0.0000	-1.3856	-.6928
Private	1.3856	0.0000	.6928
Low SES	0.0000	-1.5844	-.7319
Mid SES	-.6113	-.9731	-.1206
High SES	1.5844	0.0000	.8525

Obtain the **same odds ratios** —  
 e.g., odds ratio for public versus private

$$\text{Fix first: } \exp(0.0000 - 1.3856) = .250$$

$$\text{Fix last: } \exp(-1.3856 - 0.0000) = .250$$

$$\text{Zero sum: } \exp(-.6928 - .6928) = \exp(-1.3856) = .250$$

# Identification Constraints: example

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

Term	Dummy code		Effect code
	Fix First	Fix Last	Zero Sum
	$\beta_1 = 0$	$\beta_I = 0$	${}_i\beta = 0$
Intercept	-.8593	2.1107	.5654
Public	0.0000	-1.3856	-.6928
Private	1.3856	0.0000	.6928
Low SES	0.0000	-1.5844	-.7319
Mid SES	-.6113	-.9731	-.1206
High SES	1.5844	0.0000	.8525

Obtain the **same fitted values** —  
e.g., fitted logit for public, low SES

$$\text{logit}(\hat{\pi}) = -.8593 + 0.0000 + 0.0000 = -.8593$$

$$\text{logit}(\hat{\pi}) = 2.1107 - 1.3856 - 1.5844 = -.8593$$

$$\text{logit}(\hat{\pi}) = .5654 - .6928 - .7319 = -.8593$$

# Multiple Logistic Regression

Two or more explanatory variables where the variables may be

- Continuous
- Discrete (nominal and/or ordinal)
- Both continuous and discrete (or “mixed”).

Multiple logistic regression model as a GLM:

- **Random component** is Binomial distribution (the response variables is a dichotomous variable).
- **Systematic component** is linear predictor with more than 1 variable.

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

- **Link** is the logit

$$\text{logit}(\pi) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

# High School and Beyond Data

The response variable is whether student attend academic program

$$Y = \begin{cases} 1 & \text{if academic} \\ 0 & \text{if non-academic} \end{cases}$$

The explanatory variables are

- School type or “p” where

$$p = \begin{cases} 1 & \text{if Public} \\ 0 & \text{if Private} \end{cases}$$

- Socioeconomic status or “s” where

$$s_1 = \begin{cases} 1 & \text{if Low} \\ 0 & \text{otherwise} \end{cases} \quad s_2 = \begin{cases} 1 & \text{if Middle} \\ 0 & \text{otherwise} \end{cases}$$

We have been treating SES as a nominal variable and ignoring

- It's natural ordering
- Results from previous analyses with SES as nominal variable.

Overview

Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic Regression

# SES as a Nominal Variable

Overview

Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic Regression

Term	Fix First $\beta_1 = 0$	Fix Last $\beta_I = 0$	Zero Sum $\sum_i \beta_i = 0$	Equal Spaced Scores $s$
Low SES	0.0000	-1.5844	-.7319	1
Mid SES	0.6113	-.9731	-.1206	2
High SES	1.5844	0.0000	.8525	3

With the equally spaced scores, our model is

$$\text{logit}(\pi) = \alpha + \beta_1 p + \beta_2 s$$

Statistic	Socio-Economic Status Treated as					
	a Nominal Variable			an Ordinal Variable		
	$df$	Value	$p$ -value	$df$	Value	$p$ -value
$X^2$	2	3.748	.15	3	4.604	.20
$G^2$	2	4.623	.10	3	5.683	.13
Log Likelihood		-375.3239			-375.8542	

# SES as a Ordinal Variable

Let

$M_O$  be the model with ordinal (equal spacing) SES , and  $M_1$  be the model with nominal SES.

$M_O$  is a special case of  $M_1$ ;  $M_O$  is *nested* within  $M_1$ .

We can test whether imposing equal spacing between categories of SES leads to a significant reduction in fit using

Conditional Likelihood ratio test:

$$G^2(M_O|M_1) = G^2(M_O) - G^2(M_1) = 5.683 - 4.622 = 1.061$$

or equivalently,

$$G^2(M_O|M_1) = -2(L_0 - L_1) = -2(-375.854 - (-375.3239)) = 1.061$$

with  $df = 3 - 2 = 1$ , and  $p$ -value = .30.

**Conclusion:** Don't need unequal spaced scores; equal spacing does not lead to a significant reduction in fit of the model to the data.

# SES as a Ordinal Variable

Estimate parameters of model:

Term	Estimate	<i>ASE</i>	Wald	<i>p</i> -value
Intercept	-.3895	.379	1.05	.305
SES (s)	.7975	.129	38.26	< .0001
Public	-1.3683	.278	24.17	< .0001
Private	0.0000	—	—	—

Holding school type constant, the odds of attending an academic program are

$$\exp(.79725) = 2.22$$

times larger than the odds given an increase in SES level of 1 level or unit (i.e., from low to middle, and from middle to high).

The odds ratio for Low versus High SES equals

$$\exp(3(.79725) - 1(.79725)) = \exp(2(.79725)) = (2.22)^2 = 4.93$$

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

# SES as a Ordinal Variable

Estimate parameters of model:

Term	Estimate	<i>ASE</i>	Wald	<i>p</i> -value
Intercept	-.3895	.379	1.05	.305
SES (s)	.7975	.129	38.26	< .0001
Public	-1.3683	.278	24.17	< .0001
Private	0.0000	—	—	—

Holding SES constant, the odds of attending an academic program given public school are

$$\exp(-1.3683 - 0) = \exp(-1.3683) = .255$$

time the odds given a private school (or odds given private school is  $1/.255 = 3.93$  times larger than the odds given public school).

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

# HSB example: “Mixed” Case

- 1 nominal variable
- 1 ordinal variable
- math achievement scores. (Add this not on the basis of substantive theory, but to illustrate various concepts and techniques)

Let

$M$  = math achievement or  $x_i$  (continuous)

$S$  = SES or  $s_i$  (ordinal discrete)

$P$  = School type  $p_i$  = public or private (nominal discrete)

With these 3 variables, we'll look at

1. The possible effects of adding in additional variables on curve (relationship) between  $\pi$  and  $x$  (the continuous variable).
2. Interactions between explanatory variables in terms of modeling  $\pi$ .
3. How to select the “best” model.

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

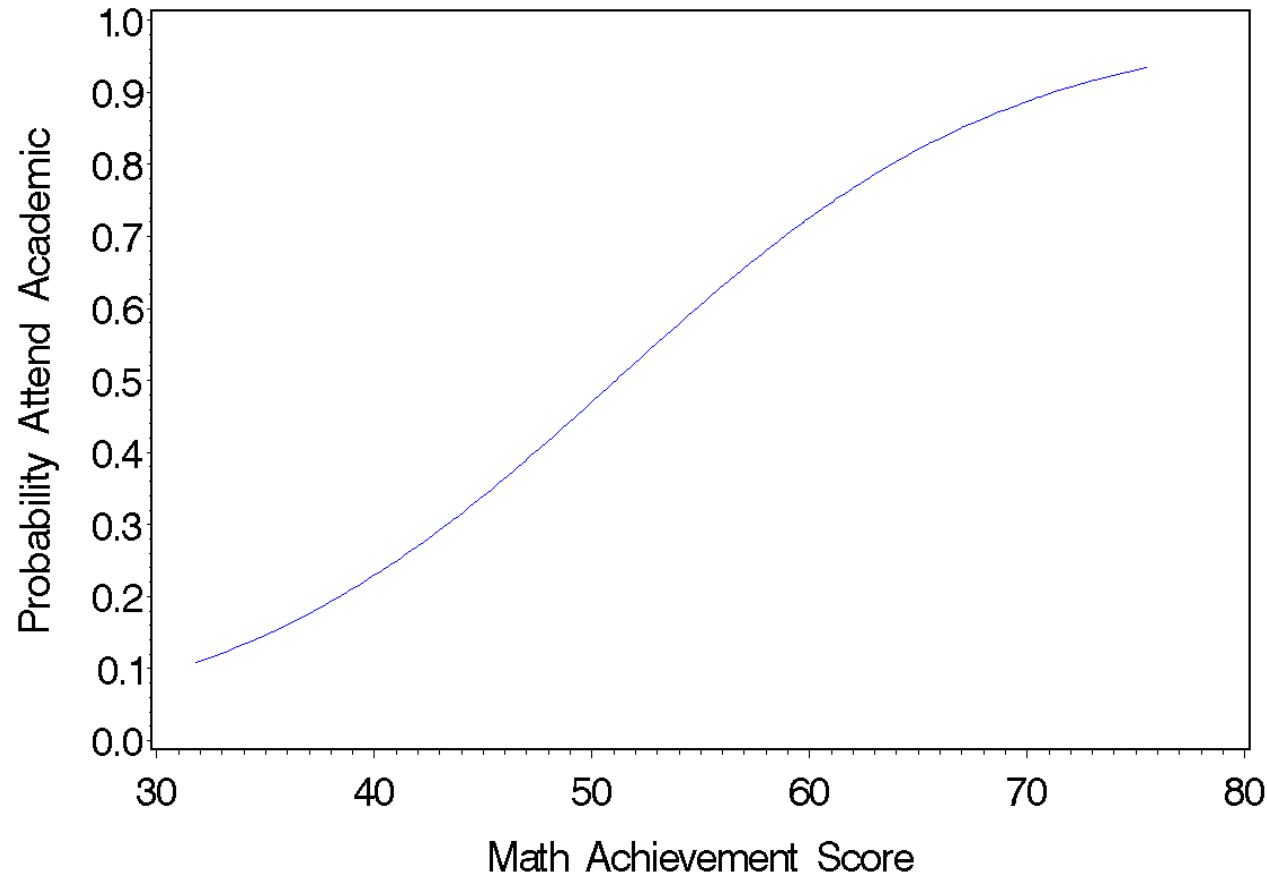
Final few Remarks on Logistic  
Regression

# Model I: just math achievement

and

$$\text{logit}(\hat{\pi}_i) = -5.5852 + 0.1093m_i$$

$$\hat{\pi}_i = \frac{\exp(5.5854 + .1093m_i)}{1 + \exp(-5.5854 + .1093m_i)}$$



Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

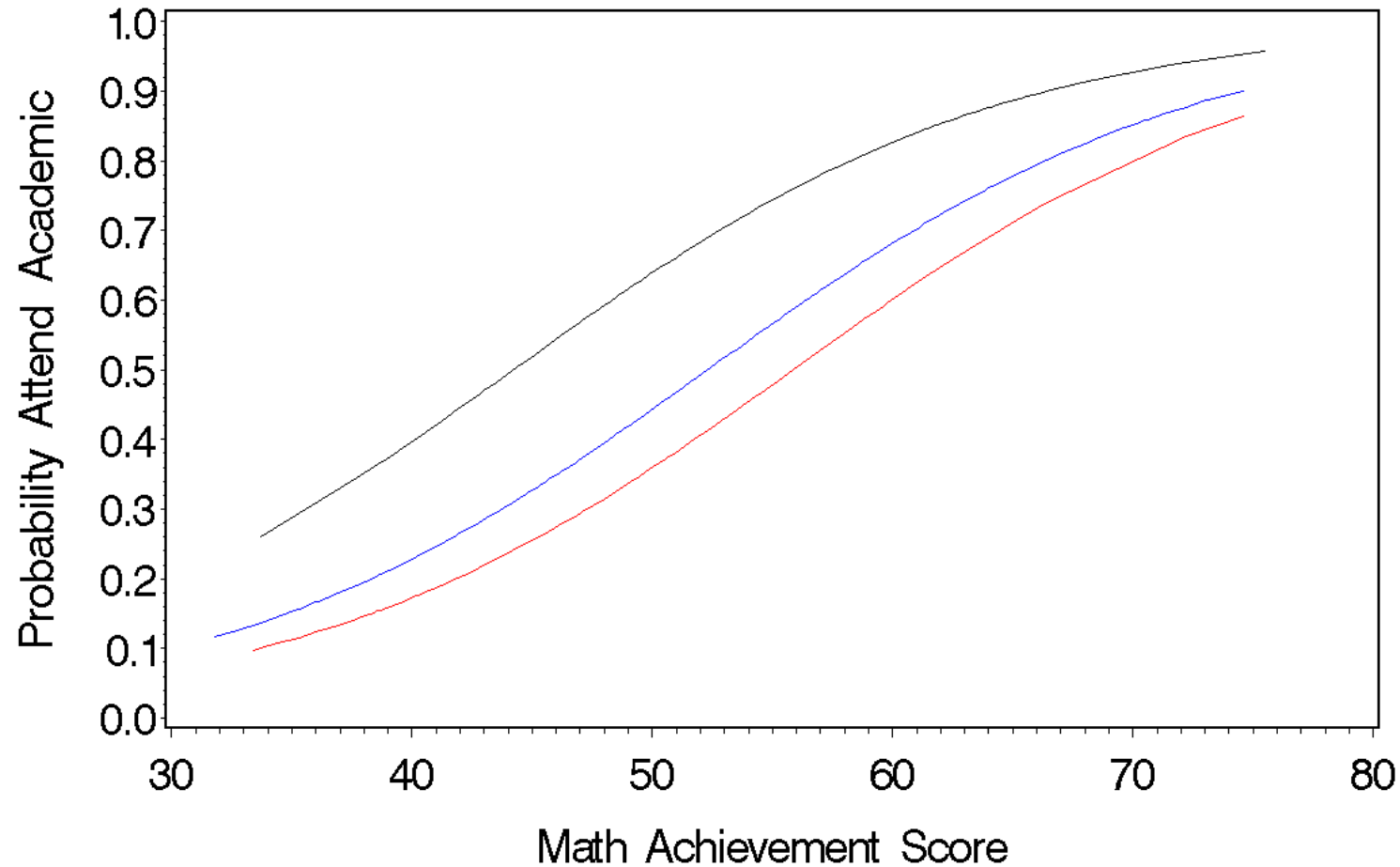
The Tale of the Titanic

Final few Remarks on Logistic  
Regression

# Model II: Add in SES as a Nominal

$$\text{logit}(\hat{\pi}_i) = -4.3733 + 0.0989m_i - 1.5003s_{1i} - 0.79966s_{2i}$$

SES: — Low — Middle — High



Overview

Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

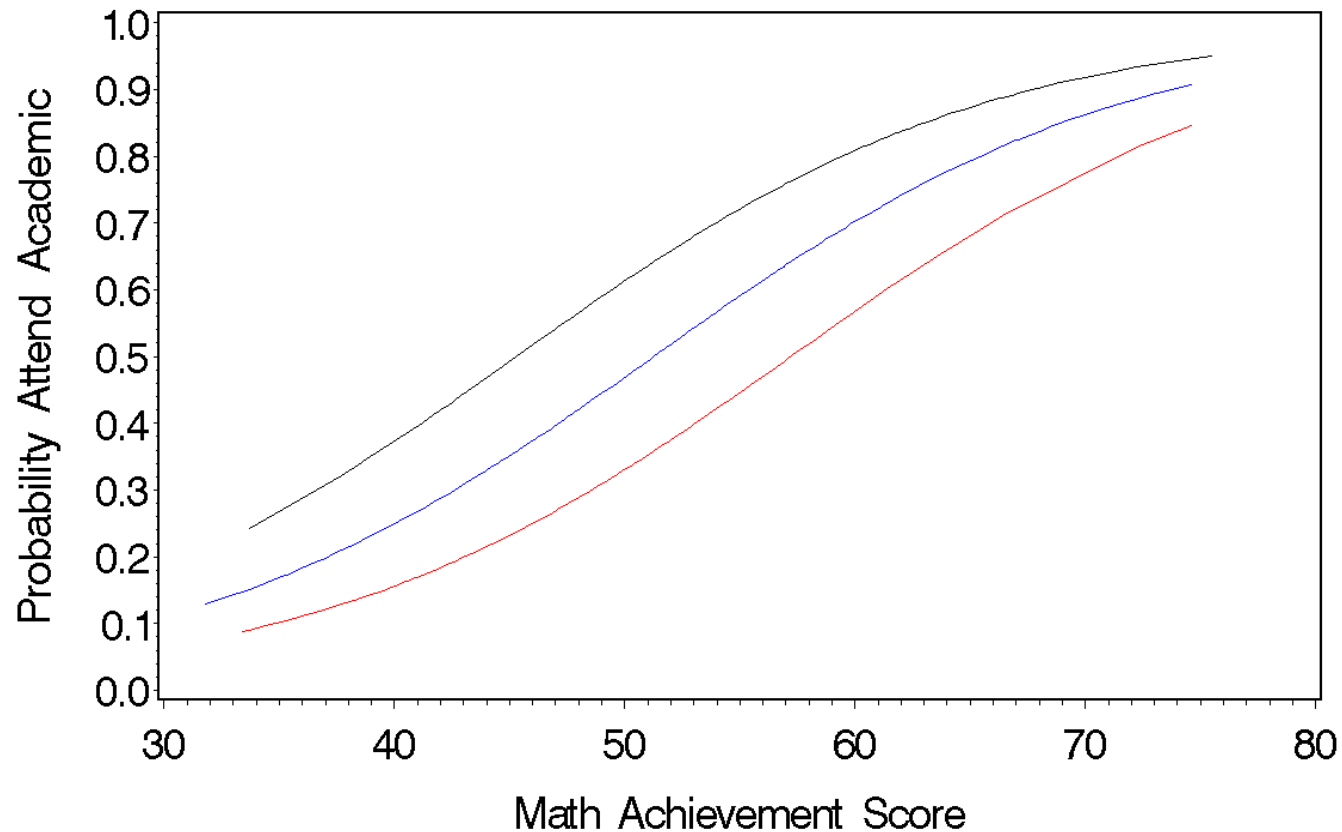
Final few Remarks on Logistic Regression

# Model III: SES as Ordinal

$$\text{logit}(\hat{\pi}_i) = -6.1914 + .0980m_i + .5837s_i$$

(Note: shape of curves the same, just equal horizontal shift.)

SES: — Low — Middle — High



Overview

Qualitative Explanatory Variables

Multiple Logistic Regression

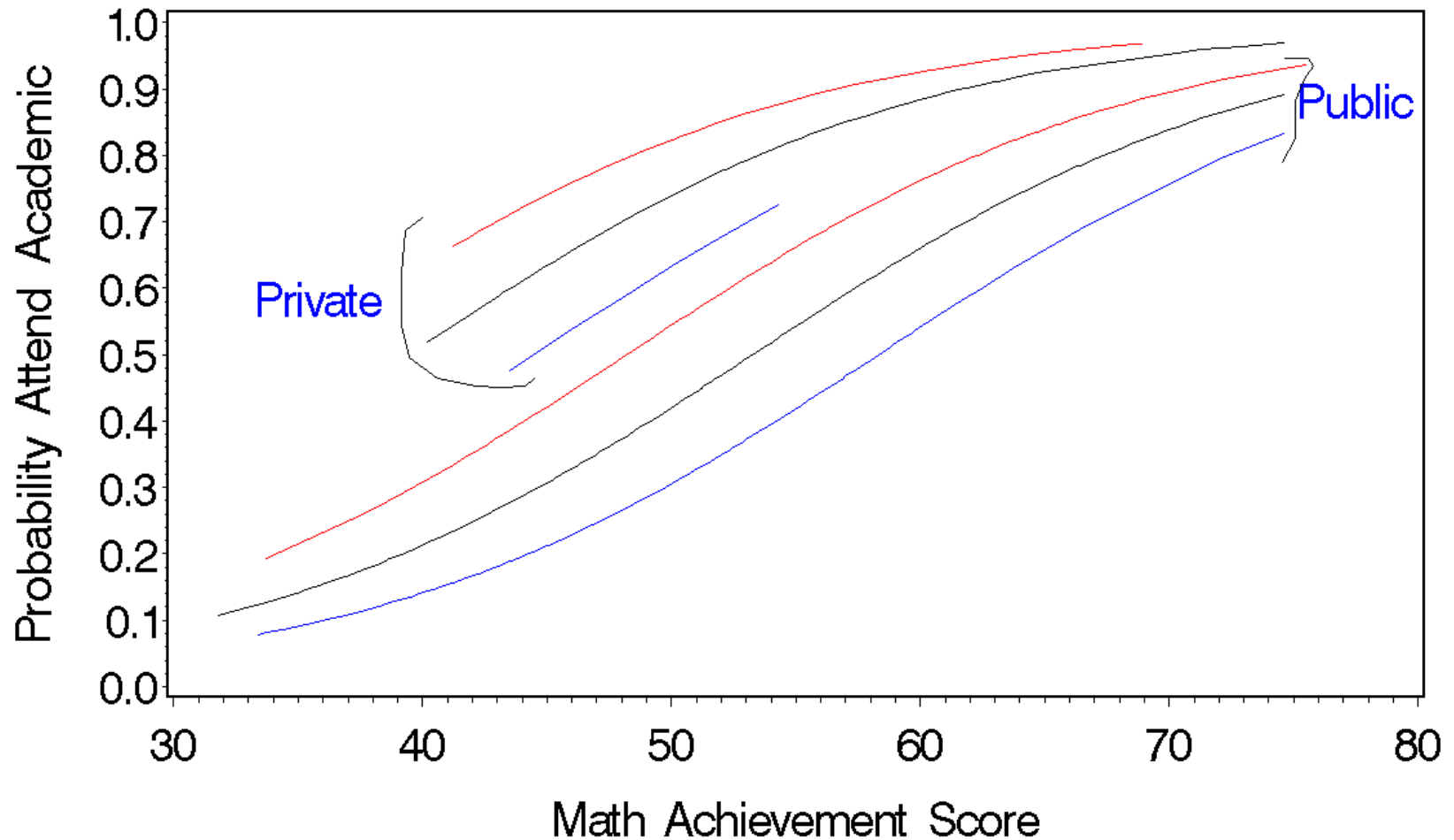
Model Selection

The Tale of the Titanic

Final few Remarks on Logistic Regression

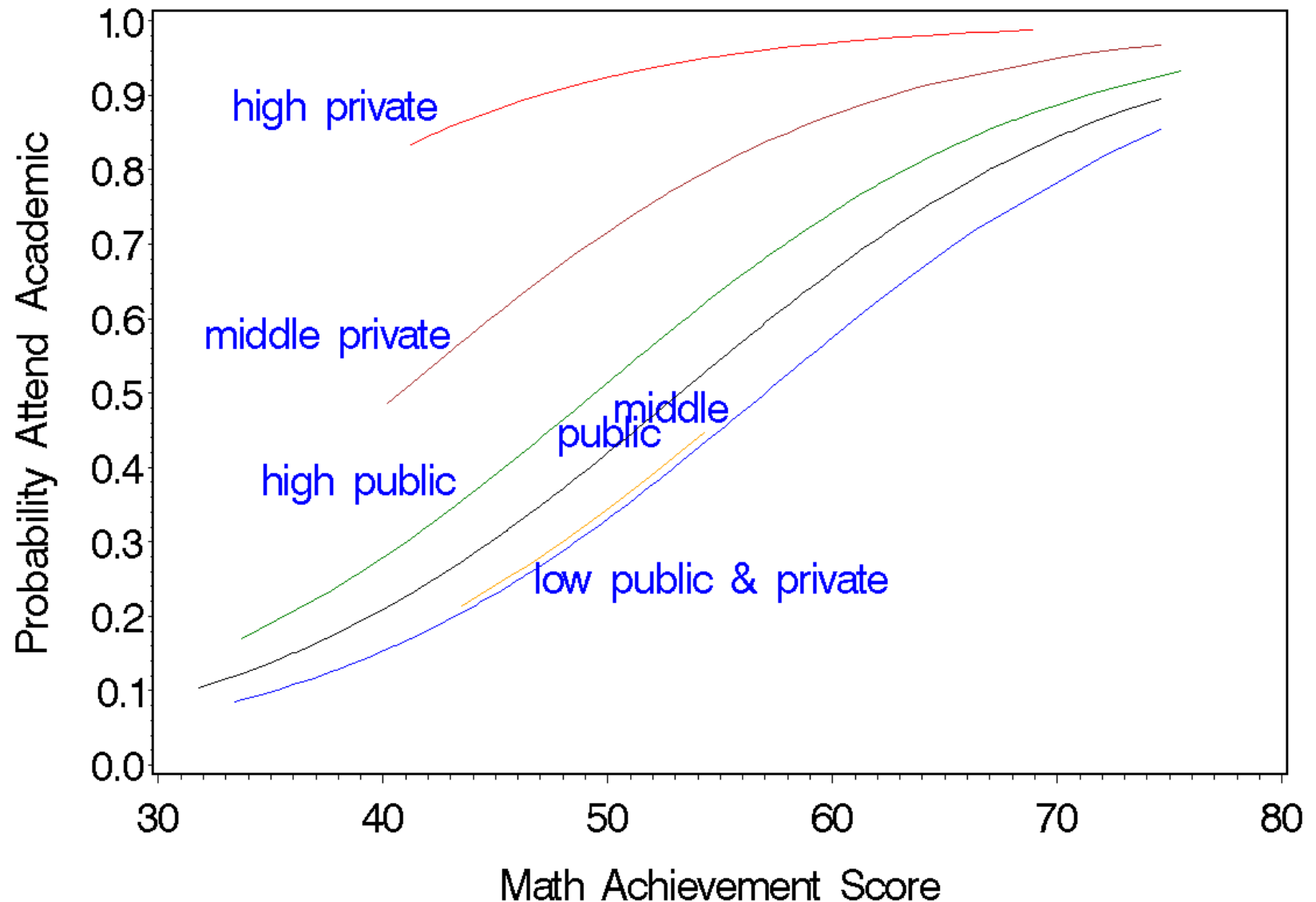
# Model IV: Add in School Type

$$\text{logit}(\hat{\pi}_i) = -5.5660 + .0986m_i + .4986s_i - 0.6823p_i$$



# Model V: Add in Interaction

$$\text{logit}(\hat{\pi}_i) = -6.6794 + .1006x_i + .9768s_i + .5663p_i - .5964(s_i p_i)$$



Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

# Model V looks pretty good

Hosmer-Lemeshow = 5.6069,  $df = 8$ ,  $p = .69$ .

Effect	DF	Wald	
		Chi-Square	Pr > ChiSq
math	1	72.6982	<.0001
ses	1	13.6467	0.0002
sctyp	1	1.0186	0.3128
ses*sctyp	1	5.0792	0.0242

Parameter	DF	Estimate	Standard	Wald	Pr > ChiSq
			Error	Chi-Square	
Intercept	1	6.6794	0.8029	69.2142	<.0001
math	1	-0.1006	0.0118	72.6982	<.0001
ses	1	-0.9768	0.2644	13.6467	0.0002
sctyp 1	1	-0.5663	0.5611	1.0186	0.3128
ses*sctyp 1	1	0.5964	0.2646	5.0792	0.0242

Adding in reading also leads to a nice model.

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

# Using Hosmer-Lemeshow Groups

Overview

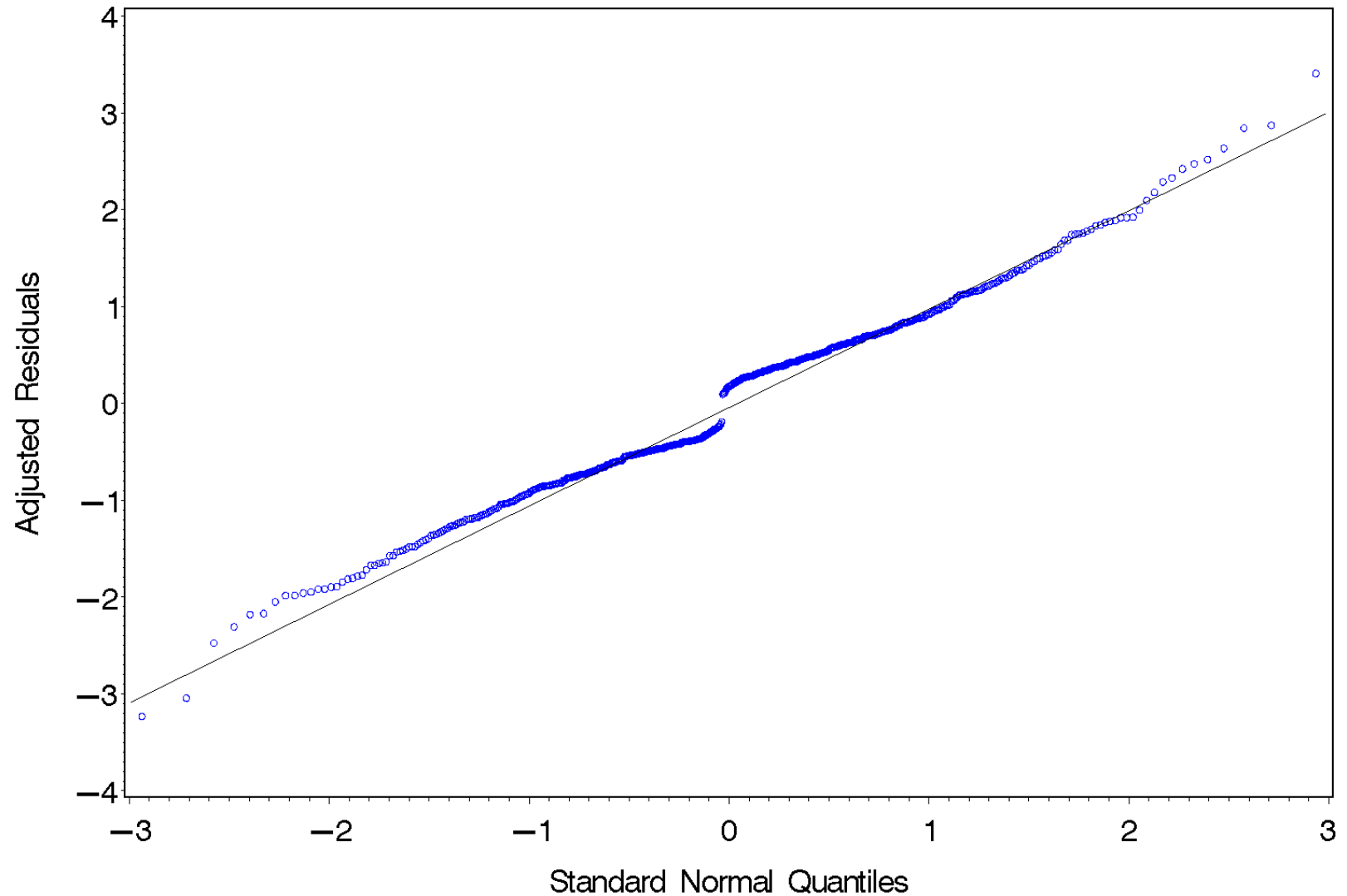
Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression



# QQ-Plot of Adjusted Residuals

Overview

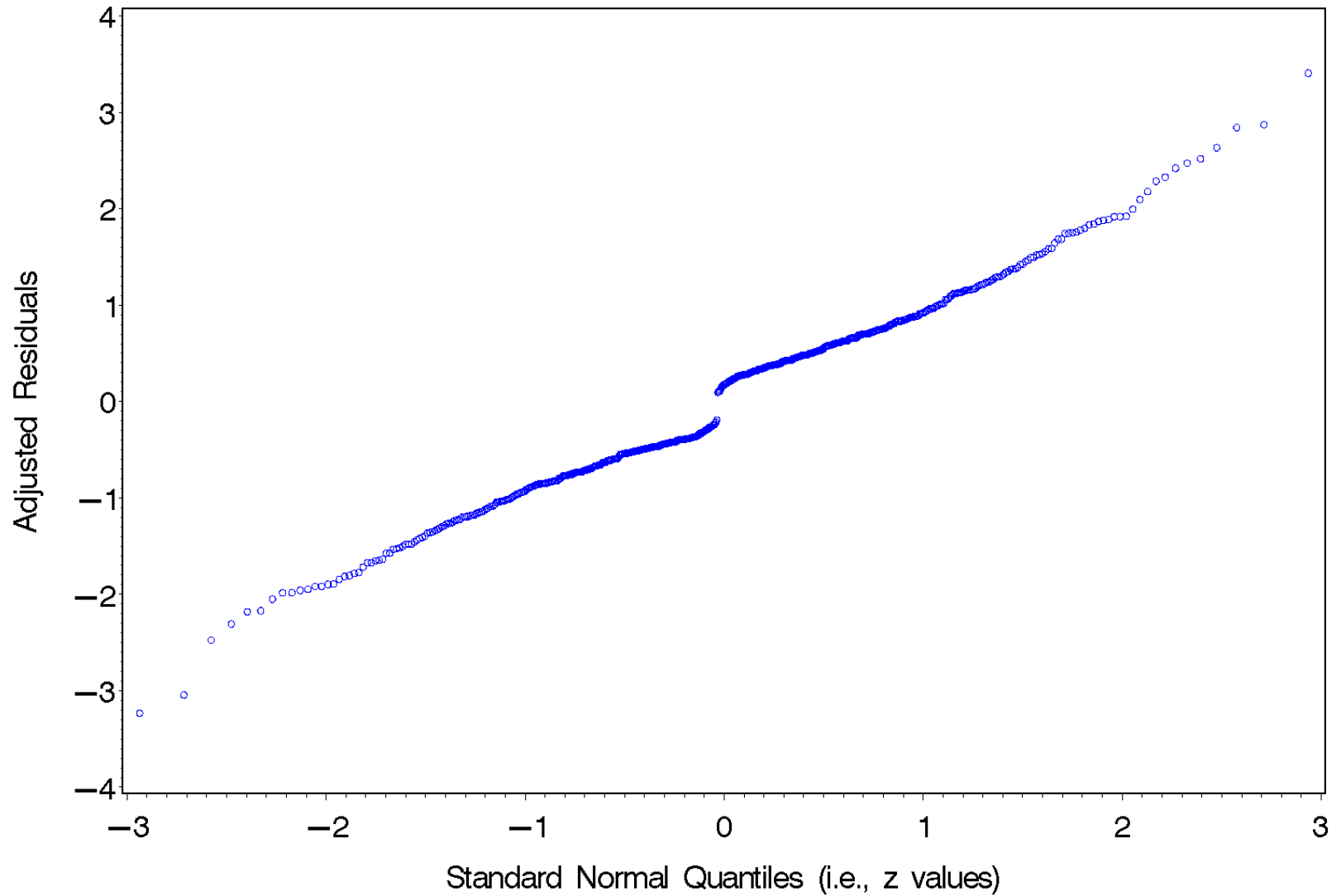
Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic Regression



# ROC for Model 5

Overview

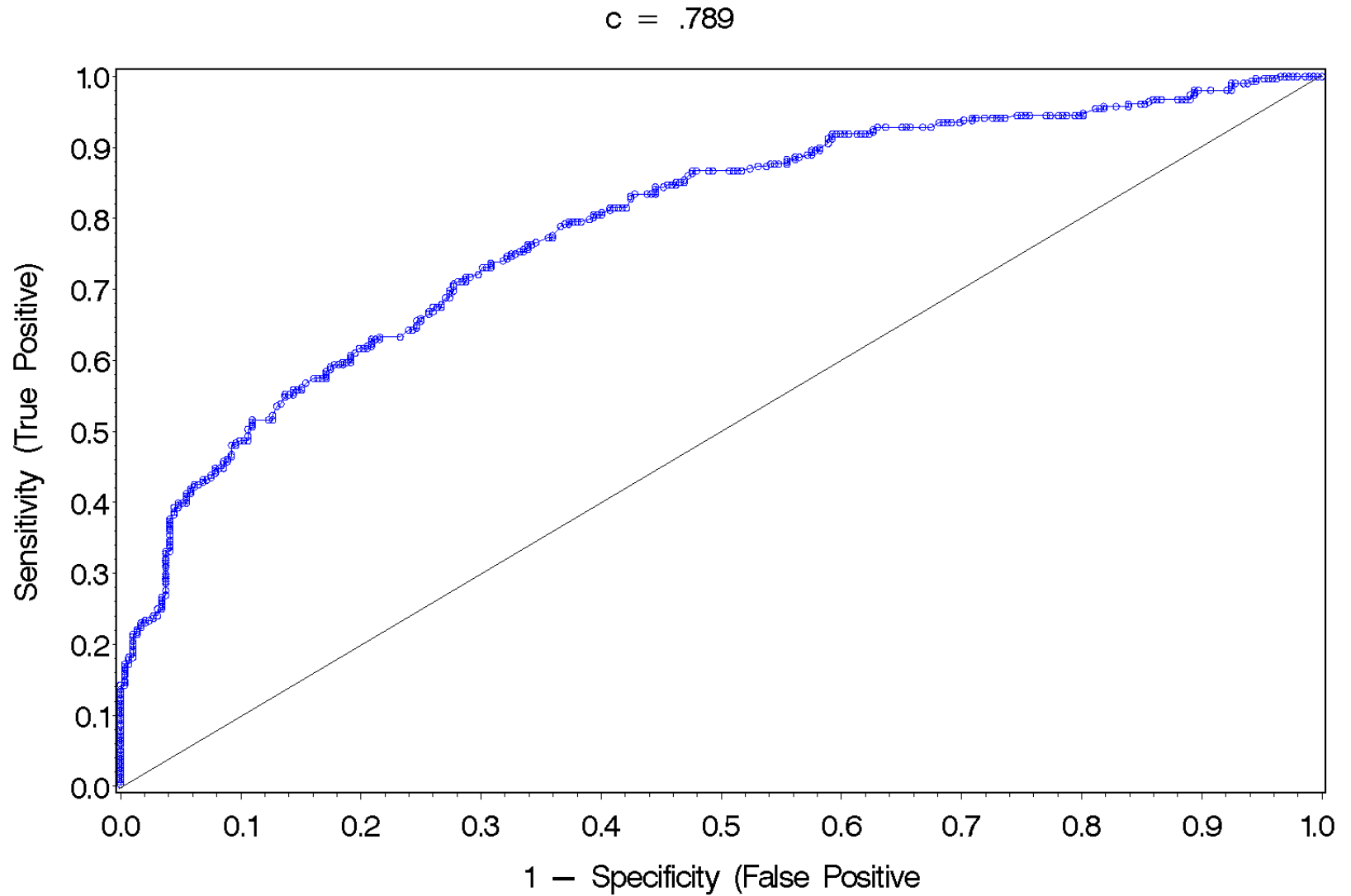
Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic Regression



# Interactions in Multiple Logistic Regression

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

1. Interaction between **two discrete variables**: the curves for  $\pi$  plotted against a continuous variables are “**shifted**” horizontally but the shape stays the same. The curves are parallel, but the distance between them need not be equal.
2. Interaction between **a continuous and a discrete variable**, then the curves would **cross**.
3. Interaction between **2 continuous variables**:
  - Plot  $\hat{\pi}$  versus values of 1 variable for selected levels of the other variable (e.g., low, middle and high value on the “other variable”).
  - If there is no interaction between the variables, the curves will be parallel.
  - If there is an interaction between the continuous variable, the curves will cross.

# HSB with more interactions

Question: What happens if we make Model 5 more complex by including other interactions?

Answer: *Nothing* is significant.

## Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard	Chi-Square	Pr>ChiSq
			Error		
Intercept	1	-6.9667	5.6284	1.5321	0.2158
MATH	1	0.1059	0.1139	0.8639	0.3526
SES	1	1.0145	2.5673	0.1562	0.6927
SCTYP 1	1	0.1458	5.6284	0.0007	0.9793
SES*SCTYP 1	1	-0.2631	2.5673	0.0105	0.9184
MATH*SES	1	-0.00051	0.0522	0.0001	0.9922
MATH*SCTYP 1	1	0.00852	0.1139	0.0056	0.9404
MATH*SES*SCTYP 1	1	-0.00664	0.0522	0.0162	0.8988

What's going on?

# Model 7

Include all five achievement variables? (writing, reading, science, civics).

Parameter	DF	Estimate	Standard	Chi-Square	Pr > ChiSq
			Error		
Intercept	1	-8.1284	0.9083	80.0802	< .0001
MATH	1	0.0664	0.0162	16.7105	< .0001
SES	1	0.9429	0.2711	12.0949	0.0005
SCTYP 1	1	0.5723	0.5674	1.0174	0.3131
SES*SCTYP 1	1	-0.6129	0.2706	5.1294	0.0235
RDG	1	0.0414	0.0156	7.0845	0.0078
SCI	1	-0.0330	0.0149	4.9040	0.0268
WRTG	1	0.0138	0.0142	0.9373	0.3330
CIV	1	0.0411	0.0132	9.7440	0.0018

Negative parameter for Science?

What's going on?

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

# Correlation Among Explanatory Variables

	RDG	WRTG	MATH	SCI	CIV	Prin1
Reading T-Score	1.0000	0.6286	0.6793	0.6907	0.5899	0.469785
Writing T-Score	0.6286	1.0000	0.6327	0.5691	0.5852	0.444455
Math T-Score	0.6793	0.6327	1.0000	0.6495	0.5342	0.457013
Science T-Score	0.6907	0.5691	0.6495	1.0000	0.5167	0.447248
Civics T-Score	0.5899	0.5852	0.5342	0.5167	1.0000	0.415777

## Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.43516810	2.90659123	0.6870	0.6870
2	0.52857686	0.11601607	0.1057	0.7927
3	0.41256079	0.08276459	0.0825	0.8753
4	0.32979620	0.03589815	0.0660	0.9412
5	0.29389805		0.0588	1.0000

Overview

Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic Regression

# Model Selection

## In Search of a Good Model

Given lots and lots of variables, which ones do we need?

### Multicollinearity

- **What:** Explanatory variables are strongly correlated; generally, one variable is about as good as another. There is redundant information in the variables.
- **Effects/Signs:**
  - ◆ Bouncing beta's.
  - ◆ If none of the Wald statistics for the variables in a model is significant, but the likelihood ratio test between the model without the variables with non-significant coefficients is significant. Rejecting the likelihood ratio test indicates that the set of variables in the model indicates that they are needed.
  - ◆ If you find that you cannot delete a variable without a significant decrease in fit but none of the parameters are significant, you might investigate whether any of the variables are correlated.

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

# Example: Chapman data ( $N = 200$ men)

**Response** is whether a person had a heart attack.

**Risk factors** considered:

- Systolic blood pressure
- Diastolic blood pressure
- Weight
- Cholesterol
- Height
- Age

Just using the three variables on the left . . .

Model	$-2 \log(L)$	Parameter	Wald	Likelihood			
		estimate	ChiSq	$p$	$df$	ratio	$p$
All 3	142.00				3	8.706	.03
systolic		-.024	1.518	.22			
diastolic		-.003	.009	.93			
weight		-.013	2.071	.15			
Just 1 explanatory variable							
Systolic	144.37	-.028	6.696	.01	1	6.339	.01
Diastolic	145.05	-.049	5.669	.02	1	6.339	.01
Weight	146.93	-.016	3.908	.05	1	3.774	.05

# Why Results so Different?

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

The correlations:

	systolic	diastolic	weight
systolic	1.00	.802	.186
diastolic		1.00	.314
weight			1.00

If you put all 6 variables in the model, age ends up being the only one that really looks significant. Age is correlated with both blood pressure measurements and weight.

# Model Selection Strategies

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

- Think. Include what you need to test your substantive research questions.
  - ◆ If you only have a few possible explanatory variables, then you could fit all possible model.
  - ◆ If you have lots and lots of variables (i.e., 4 or more), then there are various strategies that you can employ to narrow down the set of possible effects.

## ■ Backwards Elimination

1. Start with the most complex model possible (all variables and interactions).
2. Delete the highest way interaction & do a likelihood ratio test.
3. If the test is significant, stop.
4. If the test is not significant, delete each of the next highest way interaction terms & do a likelihood ratio test of the model conditioning on the model from step 2.
5. Choose the model that leads to the least decrease in fit. If the decrease in fit is not significant, try deleting highest way interactions.

6. Stop when there are no further terms that can be deleted.

# Example of Backward Elimination

With 3 explanatory variables, we could fit all possible models, but here's how the above strategy works.

$M = \text{math}$ ,  $P = \text{public (school type)}$ , &  $S = \text{SES}$

Since there is only 1 coefficient per variable, whenever we delete a term, the change in degrees of freedom will always equal 1 (i.e.,  $\Delta df = 1$ ). Therefore, we could just look at  $\Delta G^2$ . When this is not the case, you should use  $p$ -values.

	Model	$-2L$	$df$	Models Compared	$-2(L_0 - L_1)$ $= \Delta G^2$	$p$ -value	$R^2$
(1)	MSP	659.210		—			.25
(2)	MS, MP, SP	659.226		(2) — (1)	.016	.899	.24
(3a)	MS, MP	664.972		(3a) — (2)	5.746	.017	.24
(3b)	MS, SP	659.288		(3b) — (2)	.062	.803	.25
(3c)	MP, SP	659.386		(3c) — (2)	.160	.689	.25
(4a)	<b>M, SP</b>	<b>659.441</b>		(4a) — (3b)	<b>.153</b>	<b>.696</b>	<b>.25</b>
(4b)	MS, P	665.062		(4b) — (3b)	5.774	.016	.24
(5)	M, S, P	665.417		(5) — (4a)	5.977	.045	.24
(6a)	M, S	640.757					.21
(6b)	S, P	750.648					.12
(6c)	M, P	678.203					.23
(7a)	M	709.272					.18
(7b)	S	779.933					.08
(7c)	P	792.927					.06

Overview

Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic Regression

# With lots of Variables

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

- Skip the “intermediate” level models and try to “home” in on the level of complexity that you need.
- For example suppose that you have 6 possible predictors, fit
  1. Most complex model.
  2. Delete the 6–way interaction.
  3. Delete all the 5–way interactions.
  4. Delete all 4–way interaction.
  5. etc.
- What you should **NOT** do is let a computer algorithm do stepwise regression.

# Correlation Summary, $R^2$

Eight criteria used by Scott Menard (2000), Coefficient of determination for multiple logistic regression analysis. *American Statistician*, 54, 17–24.

- $R^2$  must possess utility as a measure of goodness-of-fit & have intuitively reasonable interpretation.
- $R^2$  should be dimensionless.
- $R^2$  should have well defined range and endpoints denote perfect relationship (e.g.,  $1 \leq R^2 \leq 1$  or  $0 \leq R^2 \leq 1$ ).
- $R^2$  should be general enough to apply to any type of model (e.g., random or fixed predictors).
- $R^2$  shouldn't depend on method used to fit model to data.
- $R^2$  values for different models fit to same data set are directly comparable.
- Relative values of  $R^2$  should be comparable.
- Positive and negative residuals equally weighted by  $R^2$ .

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

# Some Possible $R^2$ 's

Those reviewed by Scott Menard (2000)

## ■ OLS

$$R^2 = 1 - (SS_{error}/SS_{total}) = SS_{model}/SS_{total} = r(Y_i, \hat{Y}_i)^2,$$

- ◆ in Table, in Agresti and computed by SAS/LOGISTIC.
- ◆  $R$  is a crude index of predictive power.
- ◆ It is not necessarily decreasing as the model gets simpler.
- ◆ It depends on the range of the explanatory variables.
- ◆ It's maximum value may be less than 1 (there is a correction that SAS/LOGISTIC gives such that the maximum value is 1).

## ■ Likelihood $R^2$

- Unadjusted and Adjusted geometric mean square improvement.
- Contingency coefficient  $R^2$  and the Wald  $R^2$

There are even more proposals than those listed above.

# The Tale of the Titanic

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

- The Tale of the Titanic
- Modeling Titanic Data
- Plot using

Hosmer-Lemeshow Grouping

- QQ-plot: Pearson Residuals
- QQ-plot: Deviance Residuals
- Parameter Estimates
- Graphical Interpretation

Final few Remarks on Logistic  
Regression

- Source:
  - ◆ Thomas Carson
  - ◆ <http://stat.cmu.edu/S/Harrell/data/descriptions/titanic.html>
- Description:

The Titanic was billed as the ship that would never sink. On her maiden voyage, she set sail from Southampton to New York. On April 14th, 1912, at 11:40pm, the Titanic struck an iceberg and at 2:20 a.m. sank. Of the 2228 passengers and crew on board, only 705 survived.
- Data Available:
  - ◆  $n = 1046$
  - ◆  $Y =$  survived (0 = no, 1 = yes)
  - ◆ Explanatory variables that we'll look at:
    - Pclass = Passenger class (1 =first class, 2 =second, 3 =third)
    - Sex = Passenger gender (1 =female, 2 =male)
    - Age in years.

# Modeling Titanic Data

Another measure:  $AIC = -2\text{Log}L + 2\text{number of parameters}$   
 (smaller  $AIC \rightarrow$  the better the model)

		Models						adj	Hosmer-
Model	$-2L$	$df$	Compare	$\Delta G^2$	$p$	$AIC$	$R^2$	$R^2$	Lemshow $p$
(1) PSA	915.977	11	—	—	—	940	.3792	.5114	
(2) PS,PA,SA	917.843	9	(2)-(1)	1.866	.39	938	0.3781	.5100	.4253
(3a) PS,PA	922.174	8	(3a)-(2)	4.331	.04	940	.3755	.5065	.6357
(3b) PS,SA	927.904	7	(3b)-(1)	10.061	.01	944	.3721	.5018	.6284
(3c) PA,SA	956.004	7	(3c)-(1)	38.160	< .001	972	.3550	.4788	< .001

## Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
pclass	2	28.6170	< .0001
sex	1	19.4478	< .0001
age	1	27.5016	< .0001
pclass*sex	2	31.3793	< .0001
age*pclass	2	9.1199	.0105
age*sex	1	4.3075	.0379

Overview

Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

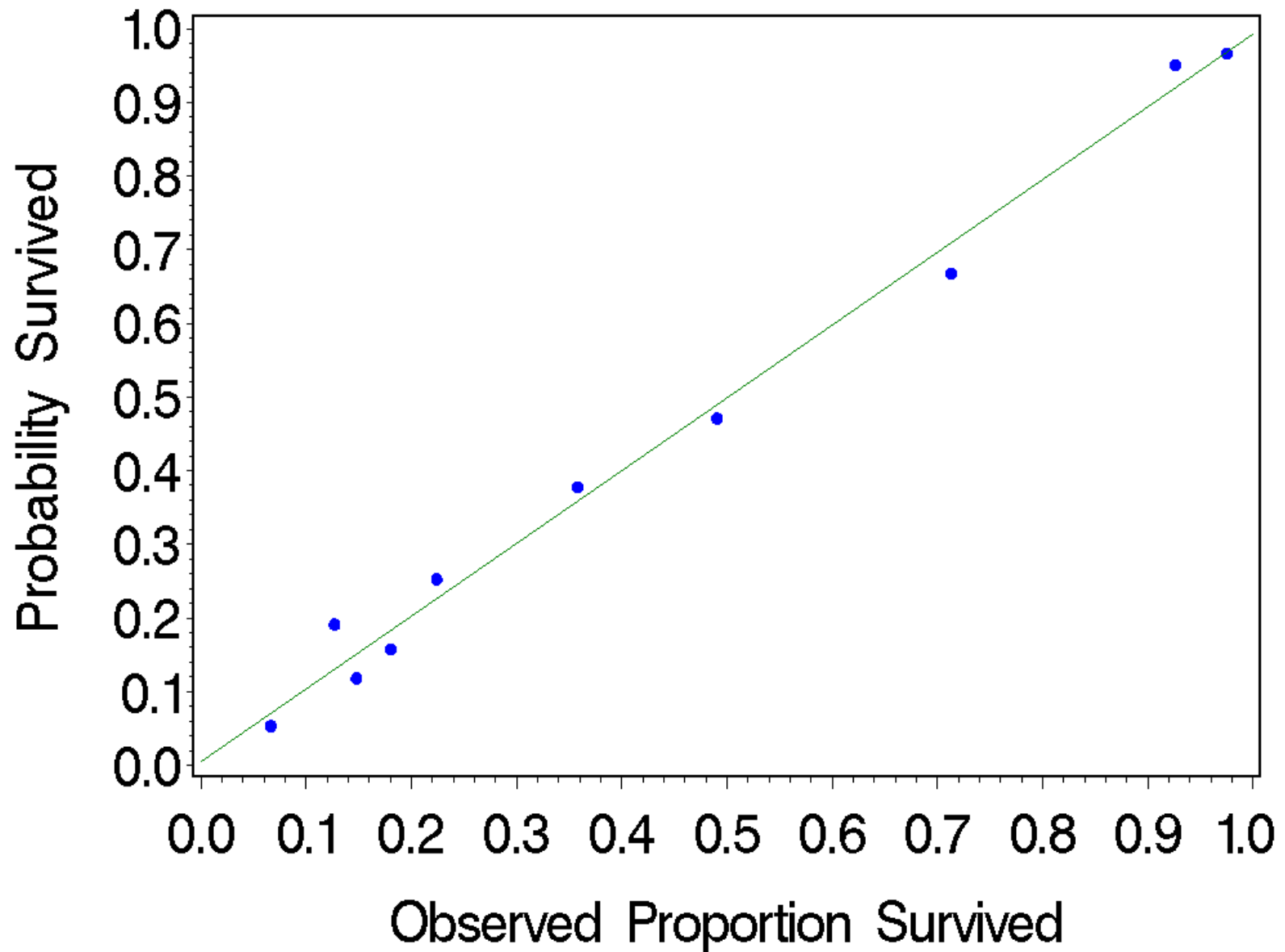
- The Tale of the Titanic
- Modeling Titanic Data
- Plot using

Hosmer-Lemeshow Grouping

- QQ-plot: Pearson Residuals
- QQ-plot: Deviance Residuals
- Parameter Estimates
- Graphical Interpretation

Final few Remarks on Logistic Regression

# Plot using Hosmer-Lemeshow Grouping



Overview

Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

● The Tale of the Titanic

● Modeling Titanic Data

● Plot using

Hosmer-Lemeshow Grouping

● QQ-plot: Pearson Residuals

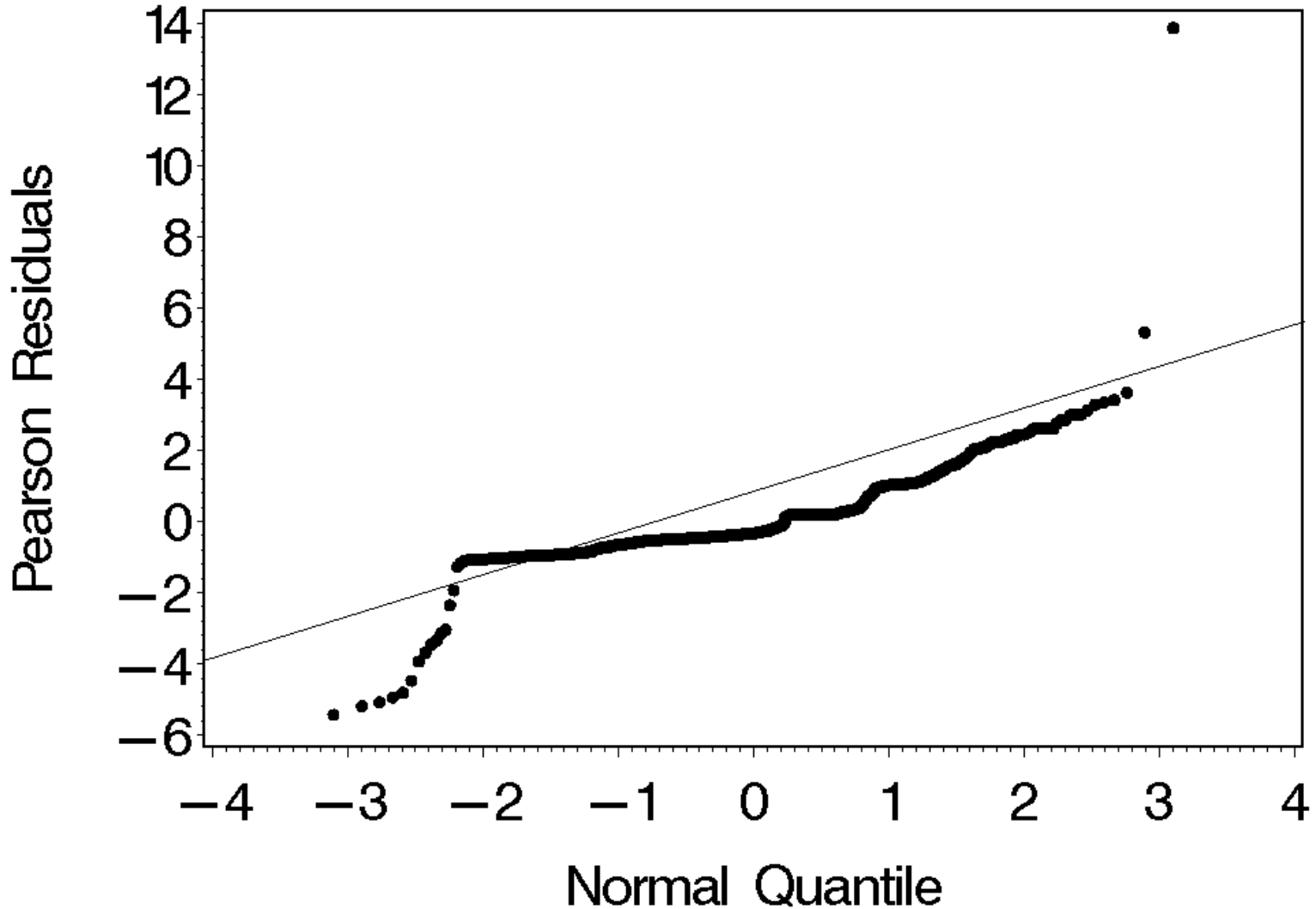
● QQ-plot: Deviance Residuals

● Parameter Estimates

● Graphical Interpretation

Final few Remarks on Logistic Regression

# QQ-plot: Pearson Residuals



Need to use SOLUTIONS>Analysis>Interactive Data Analysis

Overview

Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

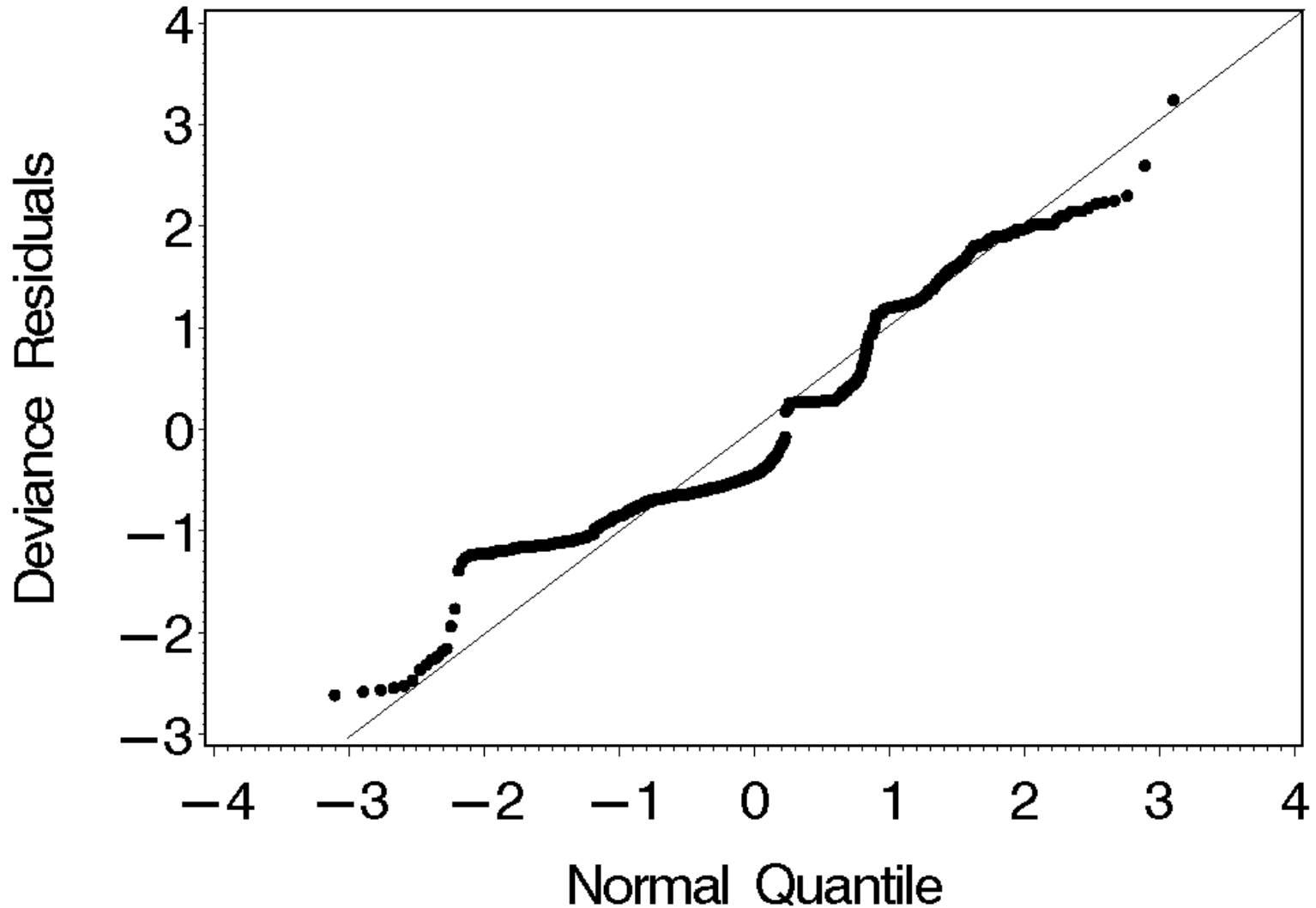
- The Tale of the Titanic
- Modeling Titanic Data
- Plot using

Hosmer-Lemeshow Grouping

- QQ-plot: Pearson Residuals
- QQ-plot: Deviance Residuals
- Parameter Estimates
- Graphical Interpretation

Final few Remarks on Logistic Regression

# QQ-plot: Deviance Residuals



Overview

Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

- The Tale of the Titanic
- Modeling Titanic Data
- Plot using

Hosmer-Lemeshow Grouping

- QQ-plot: Pearson Residuals
- QQ-plot: Deviance Residuals
- Parameter Estimates
- Graphical Interpretation

Final few Remarks on Logistic Regression

# Parameter Estimates

## Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald	Pr > ChiSq
Intercept		1	1.4269	0.2773	26.4749	< .01
pclass	1	1	0.6673	0.4104	2.6433	.10
pclass	2	1	0.9925	0.4061	5.9740	.01
sex	female	1	1.1283	0.2559	19.4478	< .01
age		1	-0.0419	0.00799	27.5016	< .01
pclass*sex	1 female	1	0.1678	0.1940	0.7480	.39
pclass*sex	2 female	1	0.6072	0.1805	11.3190	< .01
age*pclass	1	1	0.0223	0.0108	4.2606	.04
age*pclass	2	1	-0.0383	0.0127	9.1143	.00
age*sex	female	1	0.0157	0.00756	4.3075	.04

Overview

Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

● The Tale of the Titanic

● Modeling Titanic Data

● Plot using

Hosmer-Lemeshow Grouping

● QQ-plot: Pearson Residuals

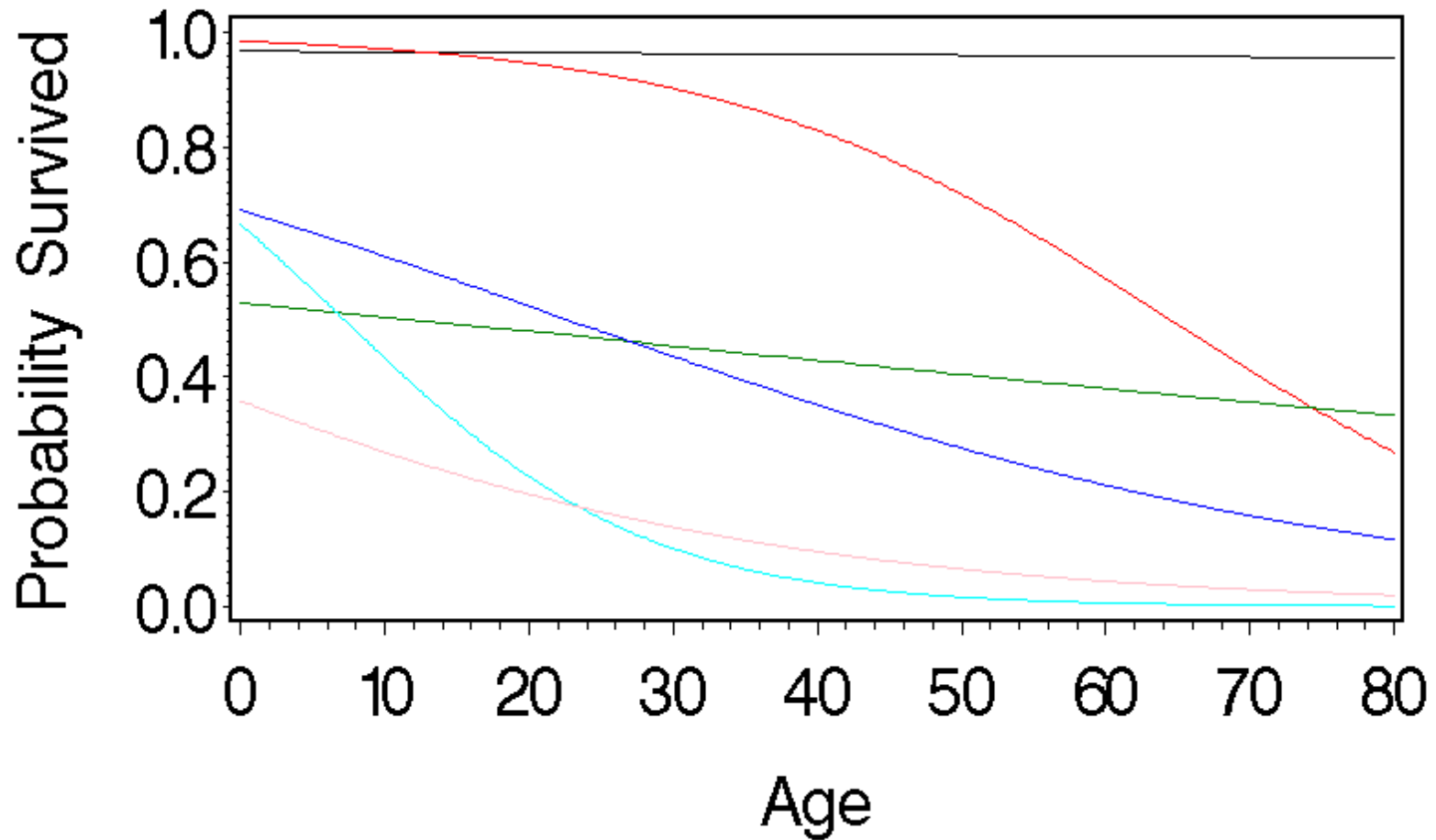
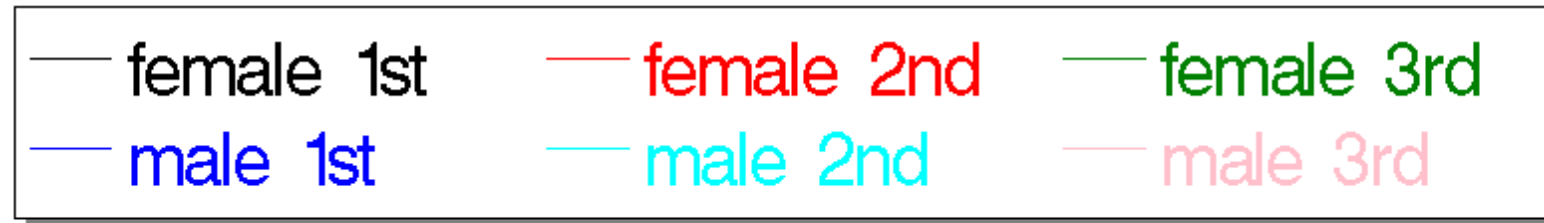
● QQ-plot: Deviance Residuals

● Parameter Estimates

● Graphical Interpretation

Final few Remarks on Logistic Regression

# Graphical Interpretation



Overview

Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

- The Tale of the Titanic
- Modeling Titanic Data
- Plot using Hosmer-Lemeshow Grouping
- QQ-plot: Pearson Residuals
- QQ-plot: Deviance Residuals
- Parameter Estimates
- Graphical Interpretation

Final few Remarks on Logistic Regression

# Final few Remarks on Logistic Regression

## Sample Size & Power.

In the text there are formulas for estimating the needed sample size to detect effects for given significance levels, power, and effect size in the case of

- One explanatory variable with 2 categories
- One Quantitative predictor.
- Multiple (quantitative) predictors.

These formulas

- give rough estimates of needed sample size.
- require guesses of probabilities, effects, etc.
- should be used at the design stage of research.

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

● Final few Remarks on Logistic  
Regression

● "Exact" Inference

● Some notes regarding SAS

● Including Interactions

● Last Two Comments on  
Logistic Regression

# “Exact” Inference

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

● Final few Remarks on Logistic  
Regression

● “Exact” Inference

● Some notes regarding SAS

● Including Interactions

● Last Two Comments on

Logistic Regression

- Maximum likelihood estimation of parameters works the best and statistics inference is valid when you have large samples.
- With small samples, you can substantially improve statistical inference by using conditional maximum likelihood estimation.
- The Basic Idea behind conditional maximum likelihood estimation:
  - ◆ Use the conditional probability distribution where you consider the “sufficient statistics” (statistics computed on data that are needed to estimate certain model parameters) as being fixed.
  - ◆ The conditional probability distribution and the maximized value of the conditional likelihood function depends only on the parameters that you’re interested in testing.
- This only works when you use the canonical link for the random component.
- The conditional method is especially useful for small samples. You can perform “exact” inference for a parameter by using conditional likelihood function that eliminates all of the other parameters).

# Some notes regarding SAS

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

● Final few Remarks on Logistic  
Regression

● "Exact" Inference

● Some notes regarding SAS

● Including Interactions

● Last Two Comments on  
Logistic Regression

- Data format is Subject  $\times$  variable matrix (1 line per subject).

- ◆ GENMOD: You need a variable for number of cases (ncases) that equals 1 for all individuals,

`model y/ncases = x1 x2 / link=logit dist=binomial;`

- ◆ LOGISTIC: You do not need number of cases,

`model y = x1 x2 / <other options you might want> ;`

- Response pattern with counts (i.e., tabular form).

- ◆ GENMOD: You need a variable for number of cases (ncases) that equals frequency count,

`model y/ncases = x1 x2 / link=logit dist=binomial;`

- ◆ LOGISTIC: You do not need number of cases (ncases) that equals a frequency count.

`model y/ncases = x1 x2 / <other options you might want> ;`

# Including Interactions

Overview

Qualitative Explanatory  
Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic  
Regression

- Final few Remarks on Logistic Regression
- "Exact" Inference
- Some notes regarding SAS
- Including Interactions
- Last Two Comments on Logistic Regression

For both **LOGISTIC** and **GENMOD**, interactions are included by using the \* notation

e.g.,

```
PROC GENMOD DATA=hsb;
```

```
CLASS public;
```

```
MODEL academic/n = math ses public ses*public
```

```
/ LINK=logit DIST=binomial;
```

Note: You need to put in all lower order effects when use use \*.

# Last Two Comments on Logistic Regression

(for now) Degrees of freedom

$$df = \text{num of logits} - \text{num of unique parameters}$$

$$= \text{num of logits} - (\# \text{parameters} - \# \text{constraints})$$

Example: High School and Beyond with school type and SES both as nominal variables. The model was

$$\text{logit}(\pi_{ij}) = \alpha + \beta_i^P + \beta_j^{SES}$$

So

$$df = (\# \text{ school types } ) \times (\# \text{ SES levels } )$$

$$- (\# \text{ unique parameters } )$$

$$= (2 \times 3) - [1 - (2 - 1) - (3 - 1)]$$

In general for a similar model with (and  $I$  and  $J$ ),

$$df = (IJ) - 1 - (I - 1) - (J - 1)$$

$$= (I - 1)(J - 1)$$

Sometimes with numerical explanatory variables, you may want to first standardize them.

Overview

Qualitative Explanatory Variables

Multiple Logistic Regression

Model Selection

The Tale of the Titanic

Final few Remarks on Logistic Regression

● Final few Remarks on Logistic Regression

● "Exact" Inference

● Some notes regarding SAS

● Including Interactions

● Last Two Comments on Logistic Regression