Multiple Logistic Regression for Dichotomous Response Variables

Edps/Psych/Soc 589

Carolyn J. Anderson

Department of Educational Psychology

ILLINOIS

Fall 2018

Outline

In last set of notes:

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

This set of notes will cover:

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

Logit models for multi-category and ordinal (polytomous) responses covered later.

Qualitative Explanatory Variables

Explanatory variables can be

- Continuous (or nearly so)
- Discrete – nominal
- Discrete – ordinal
- Continuous and Discrete (or "mixed")

We will now consider the case of discrete variables and mixed in 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 categories)
- Career choice (11 categories)
- SES level (3 levels)

HSB data yet again

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

SES Level	School Type	Program Type: non-Academic	Program Type: Academic	n_i
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

We can incorporate nominal discrete variables by creating Dummy variables (or effect codes) and include them in our model.

Dummy Variables

For School Type

$$x_1 = \begin{cases} 1 & \text{if public} \\ 0 & \text{if private} \end{cases}$$

For SES

$$s_1 = \begin{cases} 1 & \text{if low} \\ 0 & \text{otherwise} \end{cases}$$

$$s_2 = \begin{cases} 1 & \text{if middle} \\ 0 & \text{otherwise} \end{cases}$$

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., β_1) and SES (i.e., β_2 and β_3) where our dummy variables are defined as

$$\begin{aligned} x_1 &= 1 \text{ for public and } = 0 \text{ for private} \\ s_1 &= 1 \text{ for low SES and } = 0 \text{ for middle or high SES} \\ s_2 &= 1 \text{ for middle SES and } = 0 \text{ for low or high SES} \end{aligned}$$

For each combination of the explanatory variables:

SES	School Type	x_1	s_1	s_2	$\text{logit}(\pi) =$ log(academic/non-academic)
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	α

What do the parameters β_1, β_2, β_3 mean?

Interpreting β'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 given SES.

For example, for low SES,

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

Since this does not depend on an SES level (i.e., β_2 or β_3),

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

Interpreting the Other β's

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

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

Patterns of Association in 3-Way Tables

- Question: What can we say about the association in a 3-way table when the conditional odds ration 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 of a logit model 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: 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 of HSB Data

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

Statistic	df	Value	p-value
X^2	2	3.748	.15
G^2 (deviance)	2	4.622	.10
Log Likelihood		-375.324	

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

The estimated parameters, ASE and Wald statistics...

Variable/Effect	Estimate	ASE	Wald	p-value	$\exp(\beta)$
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)	$\hat{\beta}_3 = -0.9731$	.2152	20.4544	$< .001$	.38

What the Results Mean

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

- Breslow-Day Statistic $= 3.872$, $df = 2$, and $p = .14$
- CMH statistic for conditional independence of program type and school type given SES equals

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

- The conditional likelihood ratio test of the effect of school type, i.e., $H_o : \beta_1 = 0$

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

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

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

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

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

- and the one based on the logit model is

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

ANOVA-Type Representation

- When an explanatory variable has only 2 levels (e.g., school type), we only need a single dummy variable.
- When an explanatory variable 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
    - β_i^X is the parameter for the ith level of variable X.
    - β_k^Z is the parameter for the kth level of variable Z.
- Conditional independence of (say) Y and Z given Z would mean that $\beta_1^X = \beta_s^X = \ldots = \beta_I^X$.
- 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. Identification Constraints do not impact

- The estimated fitted/predicted values of π (or logit(π)); therefore, do not effect the goodness-of-fit statistics or residuals.
- The estimated odds ratios.

Impact of Identification Constraints: The constraints do effect the actual values of the parameter estimates.

The typical ID constraints are. . .

Typical Identification Constraints

Typical constraints are

- Dummy codes
  - Fix first value of a set to constant, e.g., $\beta_1 = 0$. This is what R glm does for "factors"
  - Fix the last value to a constant $\beta_I = 0$. SAS PROC GENMOD does this when use use "class".
- Effect codes
  Fix sum equal to a constant, usually 0, e.g., $\sum_{i=1}^{I} \beta_i = 0$.
  SAS PROC LOGISTIC does this.

Example of Identification Constraints

	Dummy Coded		Effect Coded
	Fix first	Fix last	Zero sum
Term	$\beta_1 = 0$	$\beta_I = 0$	$\sum_i \beta_i = 0$
Intercept	-.8593	2.1107	.5654
Public	0.0000	-1.3856	-.6928
Private	1.3856	0.0000	.6928
Low SES	0.0000	-1.5844	-.7319
Middle SES	-.6113	-.9731	-.1206
High SES	1.5844	0.0000	.8525

Obtain the same odds ratios: e.g., odds ratio of public versus private,

$$\begin{aligned}
\text{Fix first:} \quad & \exp(0.0000 - 1.3856) = \exp(-1.3856) = .250 \\
\text{Fix last:} \quad & \exp(-1.3856 - 0.0000) = \exp(-1.3856) = .250 \\
\text{Zero sum:} \quad & \exp(-.6928 - .6928) = \exp(-1.3856) = .250
\end{aligned}$$

Example continued

	Dummy Coded		Effect Coded
	Fix first	Fix last	Zero sum
Term	$\beta_1 = 0$	$\beta_I = 0$	$\sum_i \beta_i = 0$
Intercept	-.8593	2.1107	.5654
Public	0.0000	-1.3856	-.6928
Private	1.3856	0.0000	.6928
Low SES	0.0000	-1.5844	-.7319
Middle SES	-.6113	-.9731	-.1206
High SES	1.5844	0.0000	.8525

Obtain the same logit for public, low SES:

$$\begin{aligned} \text{logit}(\hat{\pi}) &= -.8593 + 0.0000 + 0.0000 = -.8593 \\ \text{logit}(\hat{\pi}) &= 2.1107 - 1.3856 - 1.5844 = -.8593 \\ \text{logit}(\hat{\pi}) &= .5654 - .6928 - .7319 = -.8503 \end{aligned}$$

Multiple Logistic Regression

Two or more explanatory variables where the variables may be

- Continuous (numerical)
- Discrete (nominal and/or ordinal)
- Both continuous and discrete (or "mixed").

Multiple logistic regression models as a GLM:

- Random component is Binomial distribution (the response variable is a dichotomous varaible).
- Systematic component is linear predictor with more than one variable:

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

- Link is the logit:

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

High School and Beyond Data

- The response variable is whether a student attended an 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

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

We have been treating SES as a nominal variable and ignoring

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

SES as Nominal Variable

Term	Fix first $\beta_1 = 0$	Fix last $\beta_I = 0$	Zero sum $\sum_i \beta_i = 0$	Equally spaced Scores, "s"
Low SES	0.0000	-1.5844	-.7319	1
Middle SES	-.6113	-.9731	-.1206	2
High SES	1.5844	0.0000	.8525	3

With the equally spaced scores we have: $\text{logit}(\pi) = \alpha + \beta_1 p + \beta_2 s$

	Socio-Economic Status Treated as a					
	Nominal Variable			Ordinal Variable		
Statistic	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 an Ordinal Variable

M_o be the model with ordinal (equal spacing here) 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 goodness-of-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,

$$G^2(M_o|M_1) = -2(L_o - L_1) = -2(-375.854 - (-.375.3239)) = 1.061$$

with $df = 3 - 2 = 1$, p-value=.30.
Conclusion: Don't need unequally spaced scores; equal spacing does not lead to a significant reduction in model fit to data.

SES as an Ordinal Variable

Estimated model parameters:

Term	Estimated	ASE	Wald	p-value
Intercept	-.3895	.379	1.05	.305
SES (s)	.7975	.129	38.26	$< .01$
Public	-1.3683	.278	24.17	$< .01$
Private	0.0000	—	—	—

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

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

times the odds given an increase in SES by 1 level (i.e., from low to middle, from middle to high).
The odds ratio for Low versus High SES equals

$$\exp(3(.79725) - 1(.79725)) = \exp(2(.79725)) = \exp(1.59459) = 4.93$$

SES as an Ordinal Variable

Estimated model parameters:

Term	Estimated	ASE	Wald	p-value
Intercept	-.3895	.379	1.05	.305
SES (s)	.7975	.129	38.26	$< .01$
Public	-1.3683	.278	24.17	$< .01$
Private	0.0000	—	—	—

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

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

times the odds given a private school. (Or the odds given private school are $1/.255 = 3.93$ times the odds for public school)

HSB Example: "Mixed" Case

- 1 nominal variable
- 1 ordinal variable
- Numerical/continuous variable

M = math achievement or x_i (continuous)
S = SES or s_i (discrete ordinal)
P = School type P_i = public or private (discrete nominal)
With these 3 variables, we'll look at

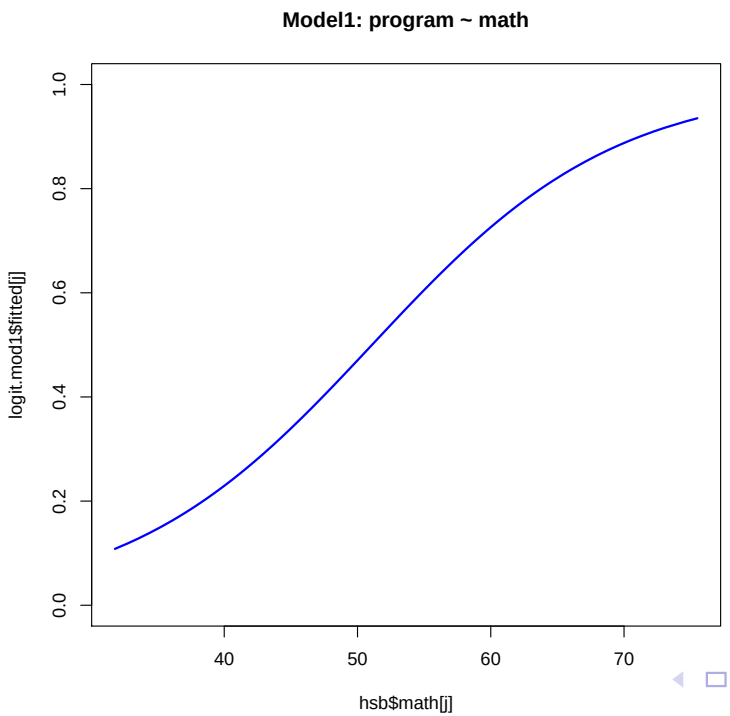
1. The possible effects of adding in additional variables on curve (relationship) between π and x (math achievement).
2. Interaction between explanatory variables in terms of modeling π.
3. How to select the "best" model.

Model I: Just math achievement

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

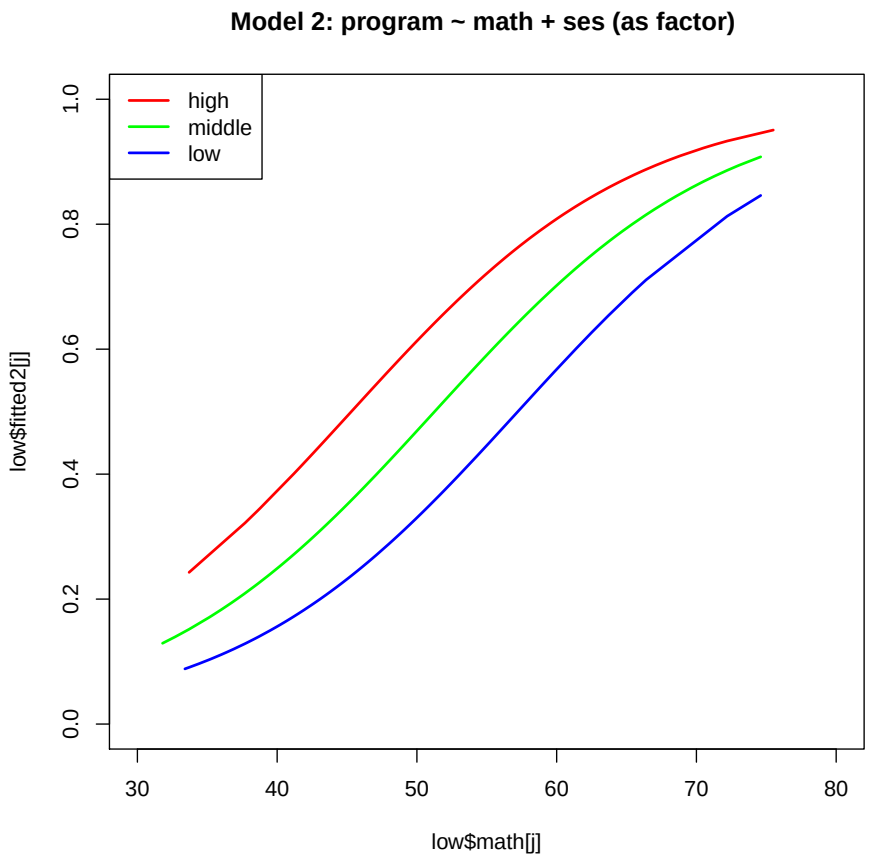
and

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



Model II: Add SES as a Nominal

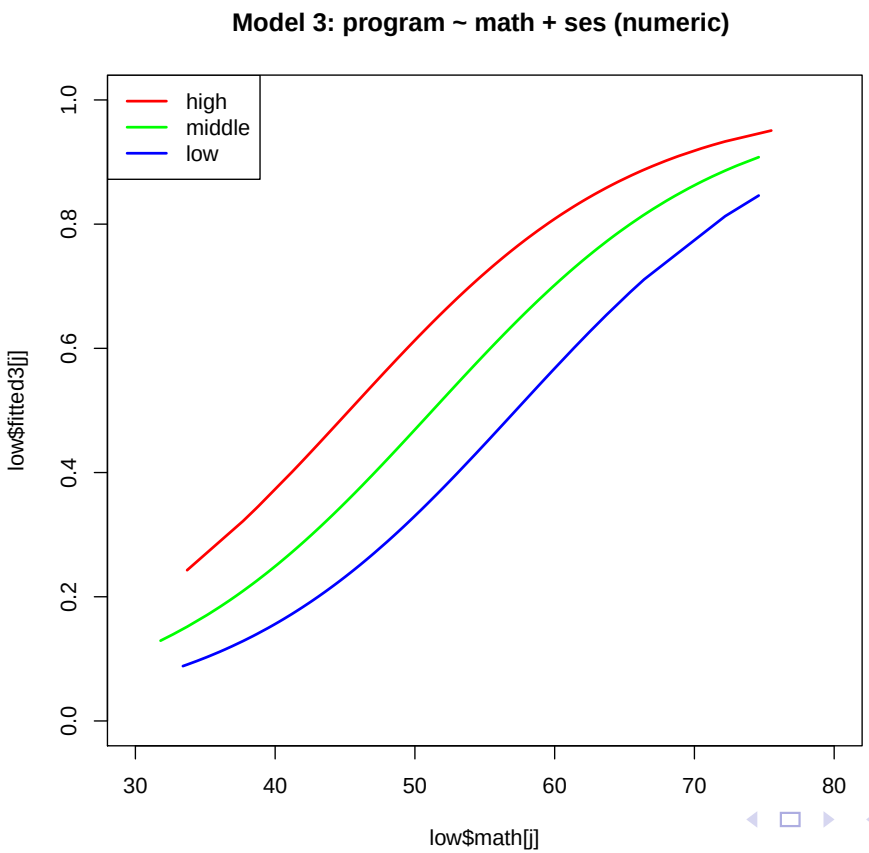
$$\text{logit}(\pi) = -4.3733 + 0.0989 m_i - 1.5003 s_{1i} - 0.79966 s_{2i}$$



Model III: SES as Ordinal

$$\text{logit}(\pi) = -6.1914 + 0.0980 m_i + 0.5837 s_i$$

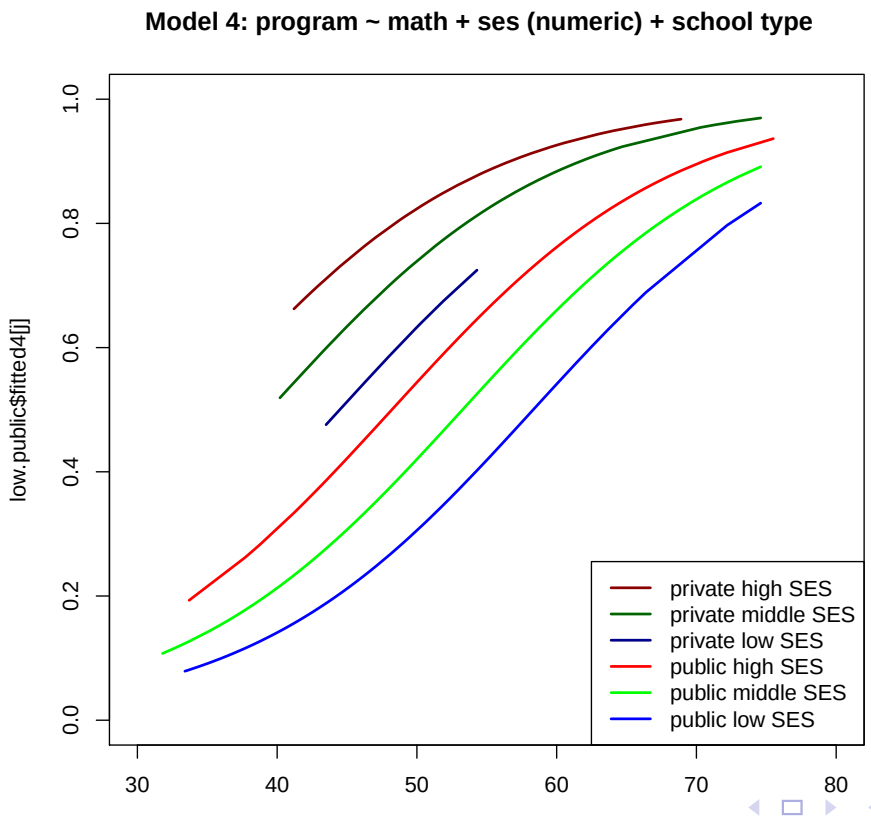
The shape of curves are same, just equal horizontal shift.



Model IV: Add School Type

$$\text{logit}(\pi) = -5.5660 + 0.0986 m_i + 0.4986 s_i - 0.6823 p_i$$

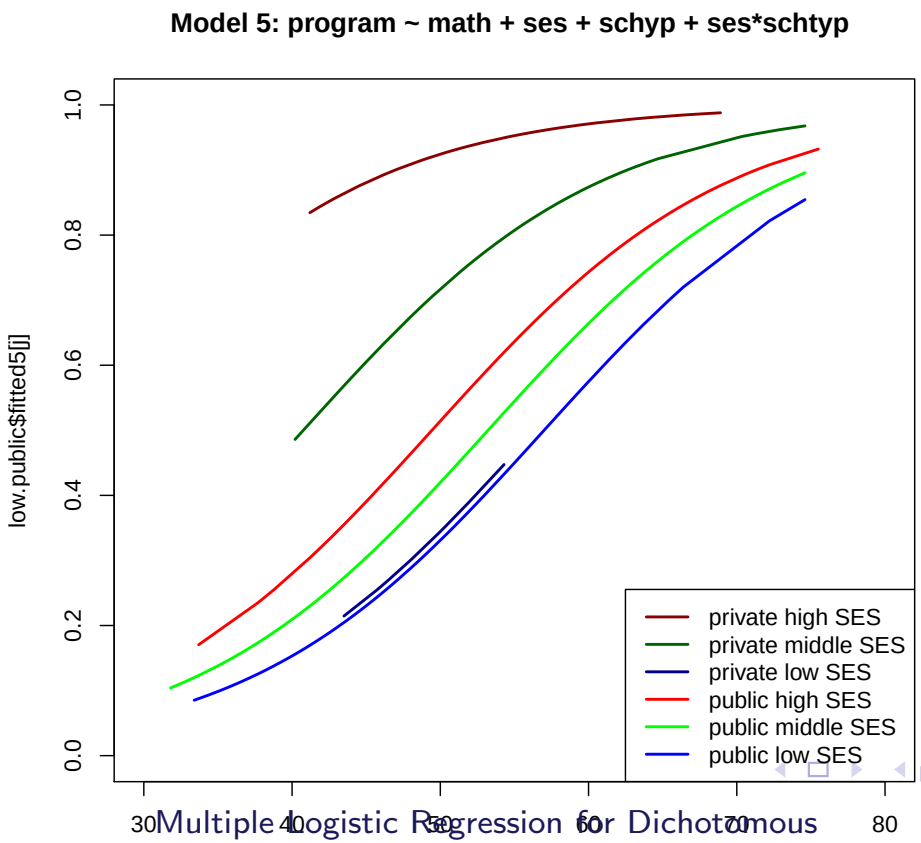
The shape of curves are same, just equal horizontal shift.



Model V: Add Interaction

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

The shape of curves are same, just equal horizontal shift.



Model V Looks Pretty Good

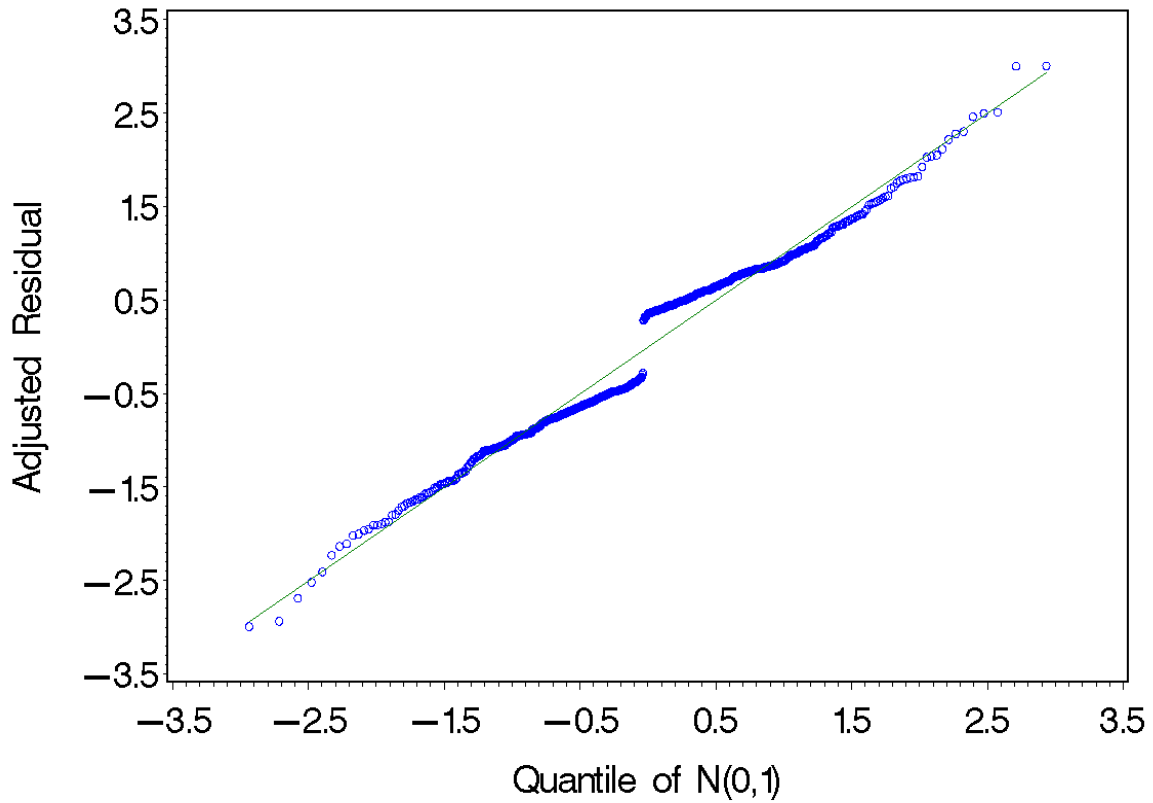
Hosmer-Lemeshow $= 5.6069$, $df = 9$, $p = .69$

Effect	df	Wald ChiSq	Pr>ChiSq
Math	1	72.6982	$< .01$
SES	1	13.6467	$< .01$
ScTyp	1	1.0186	.31
SES*SctTyp	1	5.0792	.02

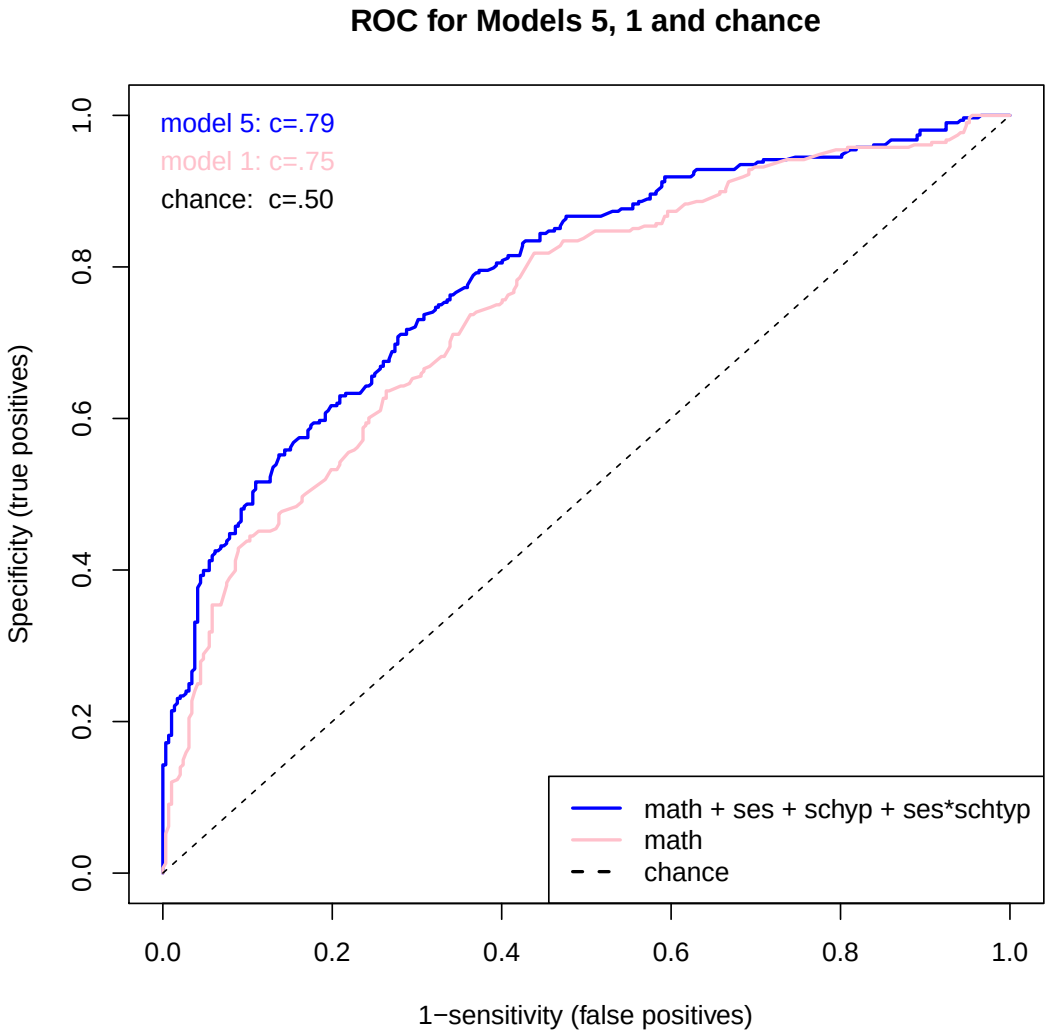
Effect	DF	Estimate	S.E.	Wald Chisq	Pr > ChiSq
Intercept	1	6.6794	0.8029	69.2142	$< .01$
Math	1	-0.1006	0.0118	72.6982	$< .01$
SES	1	-0.9768	0.2644	13.6467	$< .01$
ScTyp (public)	1	-0.5663	0.5611	1.0186	.31
SES*ScType	1	0.5964	0.2646	5.0792	.02

Adding in reading also leads to a nice model.

Model V: QQ-Plot of Adjusted Residuals



Model V: ROC Curve



Interaction in Multiple Logistic Regression

- Interaction between two discrete variables: the curves for π 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.
- Interaction between a continuous and a discrete variable will lead to curves that *cross* at some point.
- Interaction between 2 continuous variable:
  - Plot $\hat{\pi}$ versus values of one of the variables for selected levels of the other variable (e.g., 25th, 50th and 75th percentiles of the "other" variable).
  - If there is no interaction between the variables, the curves will be parallel.
  - If there is an interaction between the continuous variables, the curves will cross.

Model VI: HSB with More Interactions

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

Answer: No effects are significant! (Effect Codes)

Effect	DF	Estimate	Std Error	Chi-Square	Pr> ChiSq
Intercept	1	-6.9667	5.6284	1.5321	.22
Math	1	0.1059	0.1139	0.8639	.35
SES	1	1.0145	2.5673	0.1562	.69
ScTyp	1	0.1458	5.6284	0.0001	.98
SES*Styp	1	-0.2631	2.5673	0.0105	.92
Math*SES	1	0.0085	0.1139	0.0001	.99
Math*ScTyp	1	0.0085	0.1139	0.0056	.94
Math*SES*ScType	1	-0.0066	0.0522	0.0162	.90

What's going on?

What to do about Multicolinearity

Center the explanatory variables

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Sq	Pr > ChiSq
Intercept	1	0.5370	0.1587	11.4562	< .01
cmath	1	0.1048	0.0237	19.5444	< .01
cses	1	0.9879	0.3068	10.3693	< .01
sctyp	1	-0.6507	0.1587	16.8193	< .01
cses*sctyp	1	-0.6074	0.3068	3.9195	.05
cmath*cses	1	-0.00051	0.0522	0.0001	.99
cmath*sctyp	1	-0.00502	0.0237	0.0448	.83
cmath*cses*sctyp	1	-0.00664	0.0522	0.0162	.90

Model VII: All 5 Achievement Measures

Using effect codes (I did this in PROC LOGISTIC $\longrightarrow$ effect codes)

Effect	DF	Estimate	Std Error	Chi-Square	Pr > ChiSq
Intercept	1	-8.1284	0.9083	80.0802	< .01
math	1	0.0664	0.0162	16.7105	< .01
ses	1	0.9429	0.2711	12.0949	< .01
sctyp	1	0.5723	0.5674	1.0174	.31
ses*sctyp	1	-0.6129	0.2706	5.1294	.02
rdg	1	0.0414	0.0156	7.0845	.01
sci	1	-0.0330	0.0149	4.9040	.03
wrtg	1	0.0138	0.0142	0.9373	.33
civ	1	0.0411	0.0132	9.7440	< .01

Negative parameter for Science

What's going on?

Correlations Among Explanatory Variables

	math	rdg	sci	wrtg	civ	Prin 1
Math	1.0000	0.6793	0.6495	0.6327	0.5342	0.457013
Reading	0.6793	1.0000	0.6907	0.6286	0.5899	0.469785
Science	0.6495	0.6907	1.0000	0.5691	0.5167	0.447248
Writing	0.6327	0.6286	0.5691	1.0000	0.5852	0.444455
Civics	0.5342	0.5899	0.5167	0.5852	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

What to do?

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 the 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 deleted a variable without a significant decrease in fit but none of the estimates are significant, you might investigate whether any of the variables are correlated.

Example: Chapman Data ($N = 200$ men)

Response is whether a person had a heart attack.

Risk Factors considered:

Systolic blood pressure	Cholesterol
Diastolic blood pressure	Height
Weight	Age

Model	-2log(L)	Parameter estimate	Wald ChiSq	p	df	Likelihood ratio	p
All 3		142.00				3	8.706
	Systolic		-.024	1.518	.22		
	Diastolic		-.003	0.009	.93		
	Weight		-.013	2.071	.15		
Just 1 at-a-time							
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 are Results of Different?

The correlations between them:

	Systolic	Diastolic	Weight
Systolic	1.000	.802	.186
Diastolic	.802	1.000	.314
Weight	.186	.314	1.000

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

- Think. Include what you need to test your substantive hypotheses and to answer your research questions.
  - If you only have a few possible explanatory variables, then you could fit all possible models.
  - If you have lots and lots of variables (e.g., 4 or more), then there are various strategies that you can employ to narrow down the set of possible effects
- You can use "regularized" or "penalized" regression models (e.g., LASSO, Elastic Net). This is available in R package `glmnet`. SAS PROC GLMSELECT, but this is only OK for normal linear regression.
- Backwards elimination.

Backwards Elimination

1. Start with the most complex model possible (all variables and all 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-ways 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 the model goodness-of-fit. If the decrease in fit is not significant, try deleting the highest way interactions.
6. Stop when there are no further terms that can be deleted.

Example of Backwards Elimination

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

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

Since only 1 coefficient per variable, the change in degrees of freedom will always equal 1; therefore, we could just look at ΔG^2. When this is not the case, you should use p-values.

	Model	$-2L$	df	Compared	ΔG^2	p-value	R^2
(1)	MSP	659.210		—	—		.25
(2)	MS, MP, SP	659.226		(2)-(1)	0.916	.899	.24
(3a)	MS, MP	664.972		(3a)-(2)	5.746	.017	.24
(3b)	ME, SP	659.288		(3b)-(2)	0.062	.803	.25
(3c)	MP, SP	659.386		(3c)-(2)	0.160	.689	.25
(4a)	M, SP	659.441		(4a)-(3c)	0.153	.695	.25
(4b)	MS, P	665.062		(4b)-(3c)	6.774	.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

With Lots of Variables

- Skip the "intermediate" level models and try to hone in on the level of complexity that is needed.
- For example, suppose that you have all 6 possible predictors, fit

  1. Most complex model.
  2. Delete the 6-way interaction.
  3. Delete all of the 5-way interactions.
  4. Delete all of the 4-way interactions
  5. etc.

- What you should NOT do is let a computer algorithm do the stepwise regression.

Correlation Summary, R^2

There are at least 8 different ones. Eight Criteria summarized 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 and have intuitively reasonable interpretation.
- R^2 should be dimensionless.
- R^2 should have well defined range and end points 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 should not depend on method used to fit model to data.
- R^2 values for different models fit to the same data set are directly comparable.
- Relative value of R^2 should be comparable
- Positive and negative residuals are equally weighted by R^2.

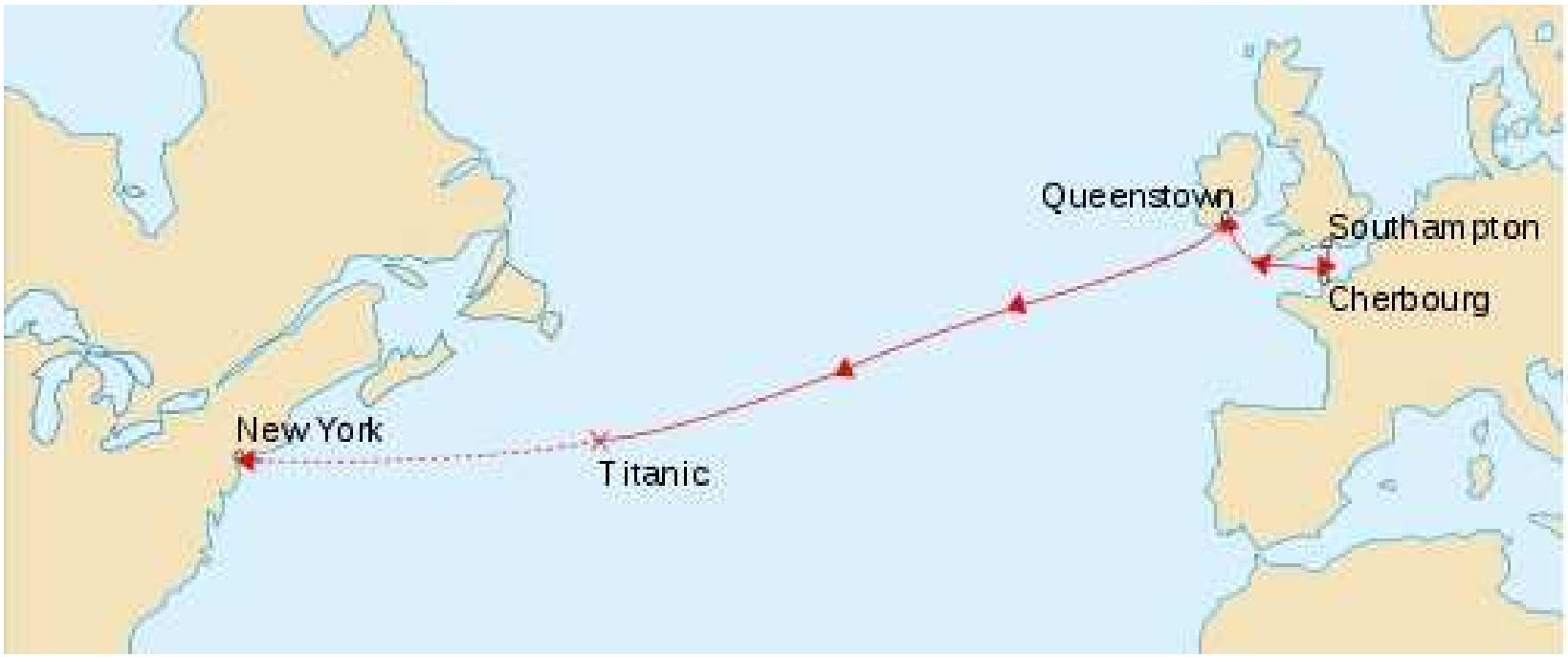
Some Possible R^2

- OLS (ordianary least squares)

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

  - In Table on page 41 & Agresti (from PROC LOGISTIC).
  - R^2 is a crude index of predictive power.
  - It is no 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 (PROC LOGISTIC has a correction such that maximum can be 1).
- Likelihood R^2
- Unadjusted and adjusted geometric mean square improvment.
- Contingency coefficient R^2 and the Wald R^2.
- and more...

The Tale of the Titanic



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.

Titanic Data Set

The data can be found on course web-site and online
For more information, goggle "Titanic data set"

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 class, 3 =third class)
  - Sex = Passenger gender (1 =female, 2 =male)
  - Age in years.

I used SAS PROC LOGISTIC, i.e., effect coding, but the default dummy coding in R glm.

Modeling the Titanic Data Set

Another measure:

$AIC = -2\log(\text{Likelihood}) - 2(\text{number of parameters})$

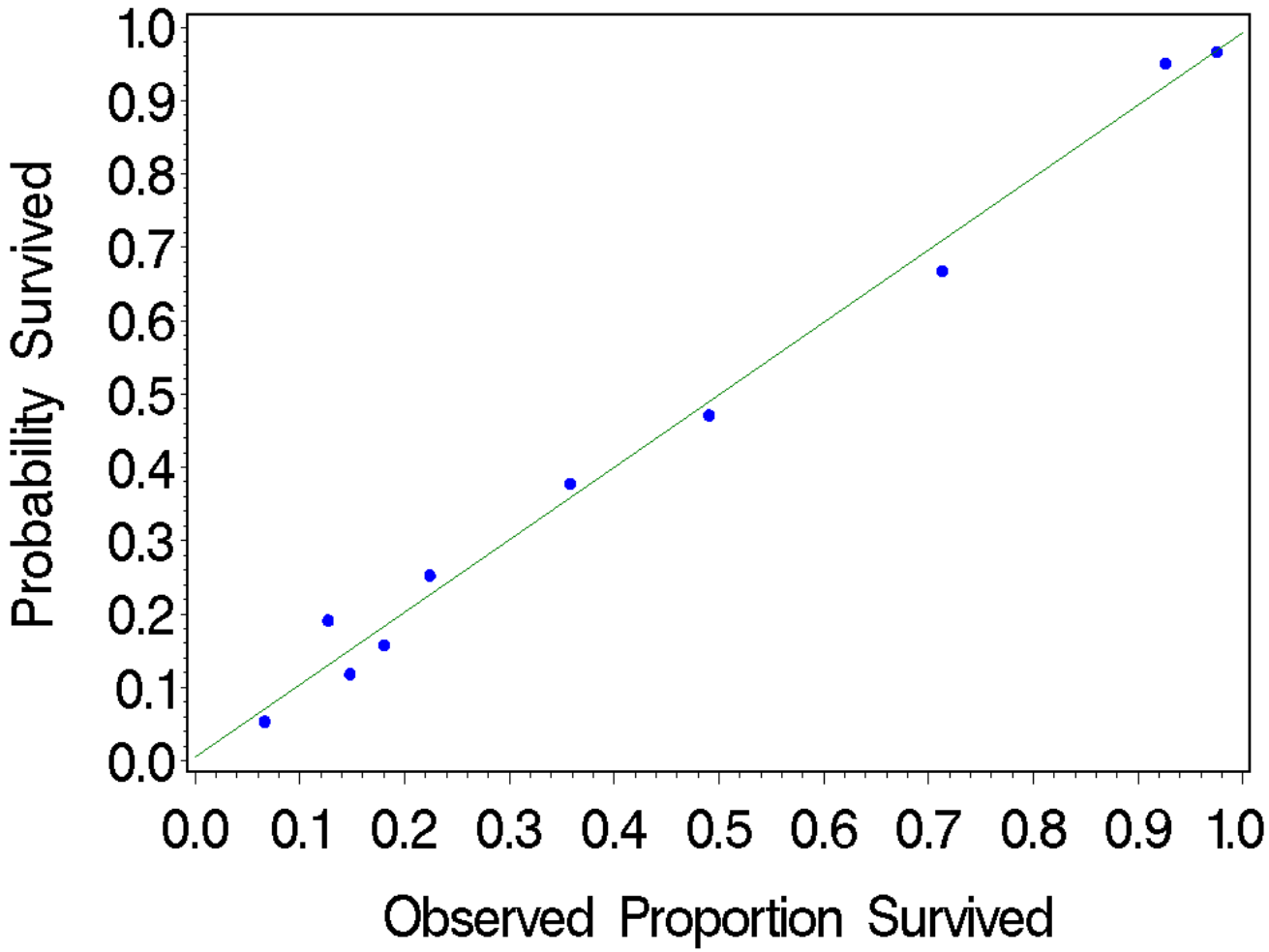
The smaller AIC $\longrightarrow$ the better the model.

Model	$-2L$	df	ΔG^2	p	AIC	R^2	adj R^2	Hosmer–Lemshow
PSA	915.977	11	—	—	940	.38	.51	
PS,PA,SA	917.843	9	1.866	.39	938	.38	.51	.43
PS,PA	922.174	8	4.331	.04	940	.38	.51	.64
PS,SA	927.904	7	10.061	.01	944	.37	.50	.63
PA,SA	956.004	7	38.160	$< .01$	972	.36	.48	$< .01$

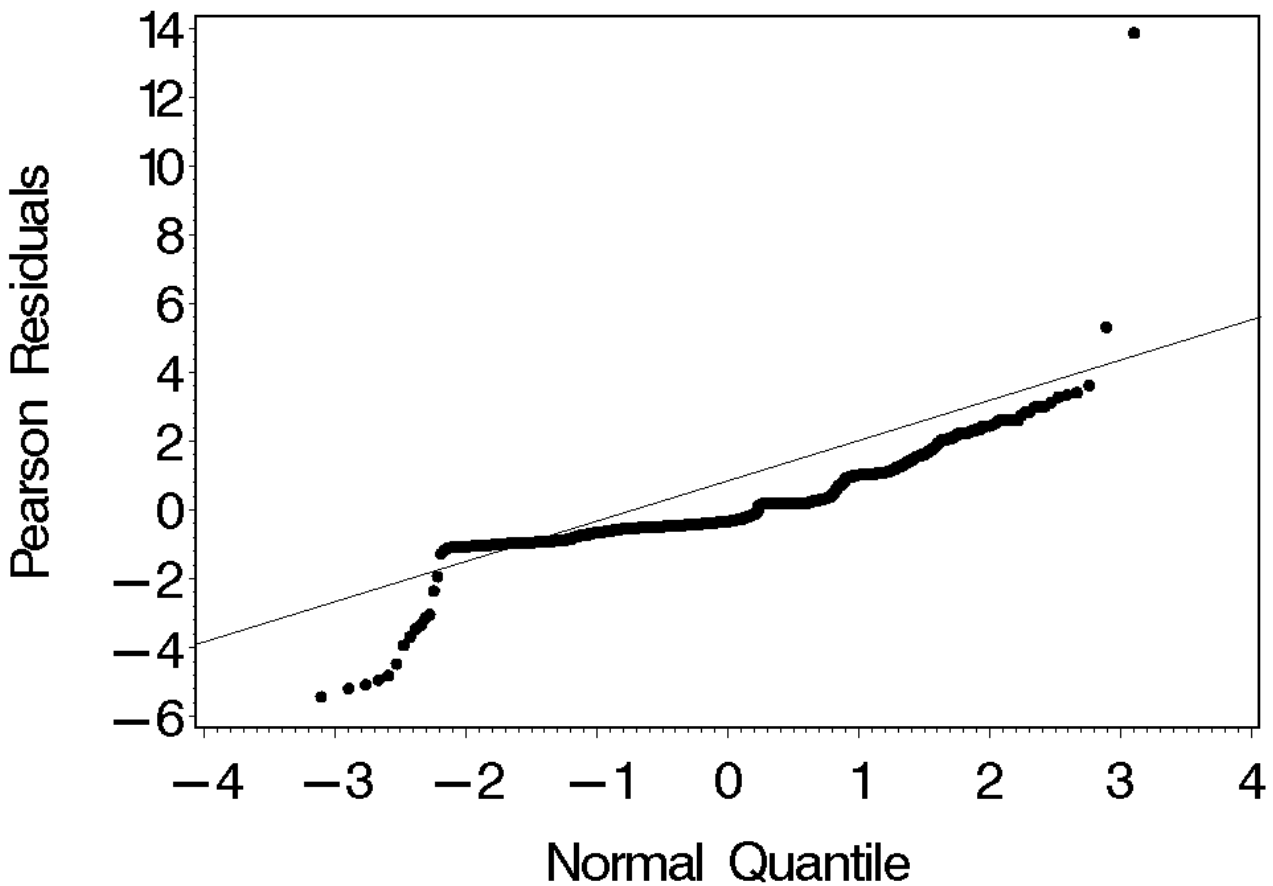
Type 3 Analysis of Effects

Effect	df	Wald Chi-square	pr >chi-square
pclass	2	28.6170	$< .01$
sex	1	19.4478	$< .01$
age	1	27.5016	$< .01$
pclass*sex	2	31.3793	$< .01$
age*pclass	2	9.1199	.01
age*sex	1	4.3075	.04

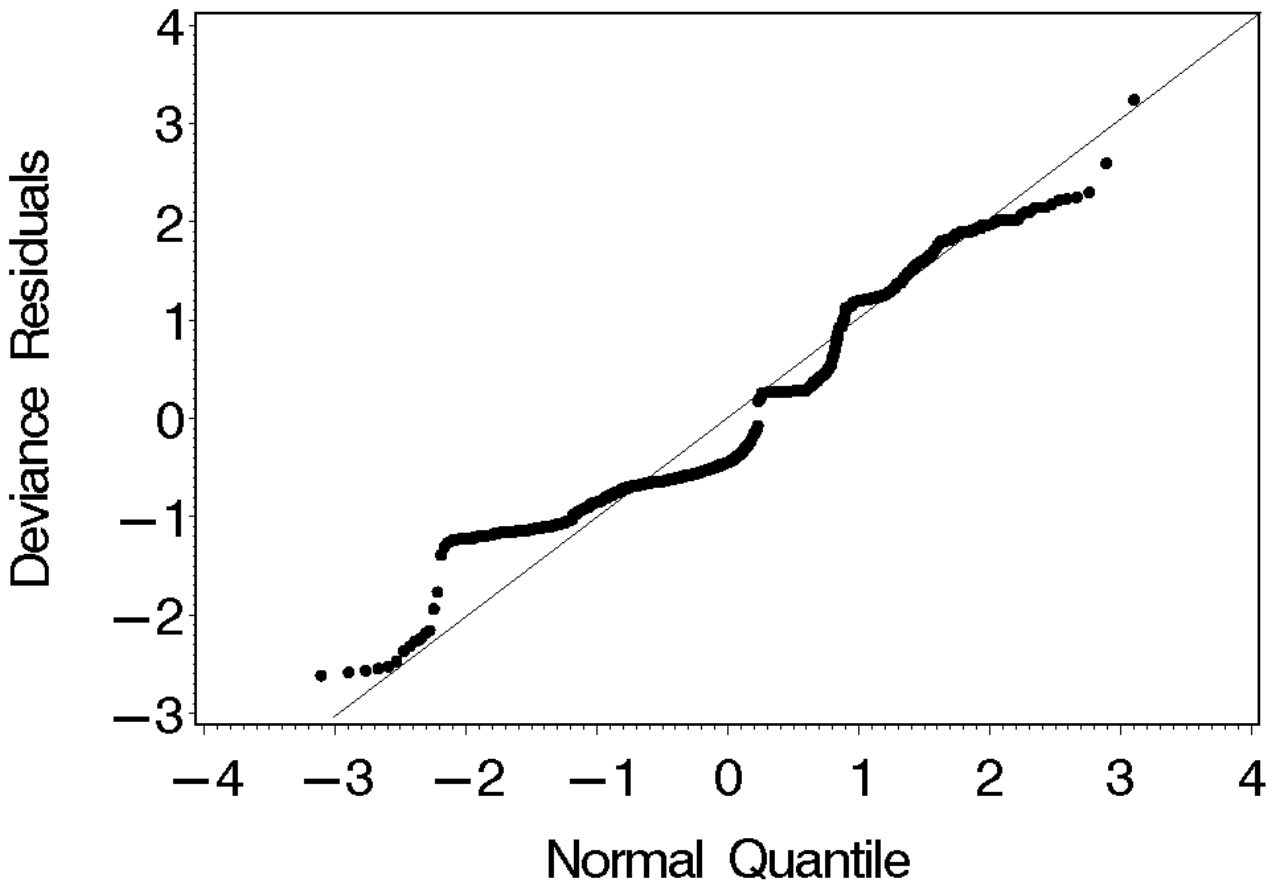
Using Hosmer-Lemshow Grouping



QQ-Plots of Pearson Residuals



QQ-Plots of Deviance Residuals

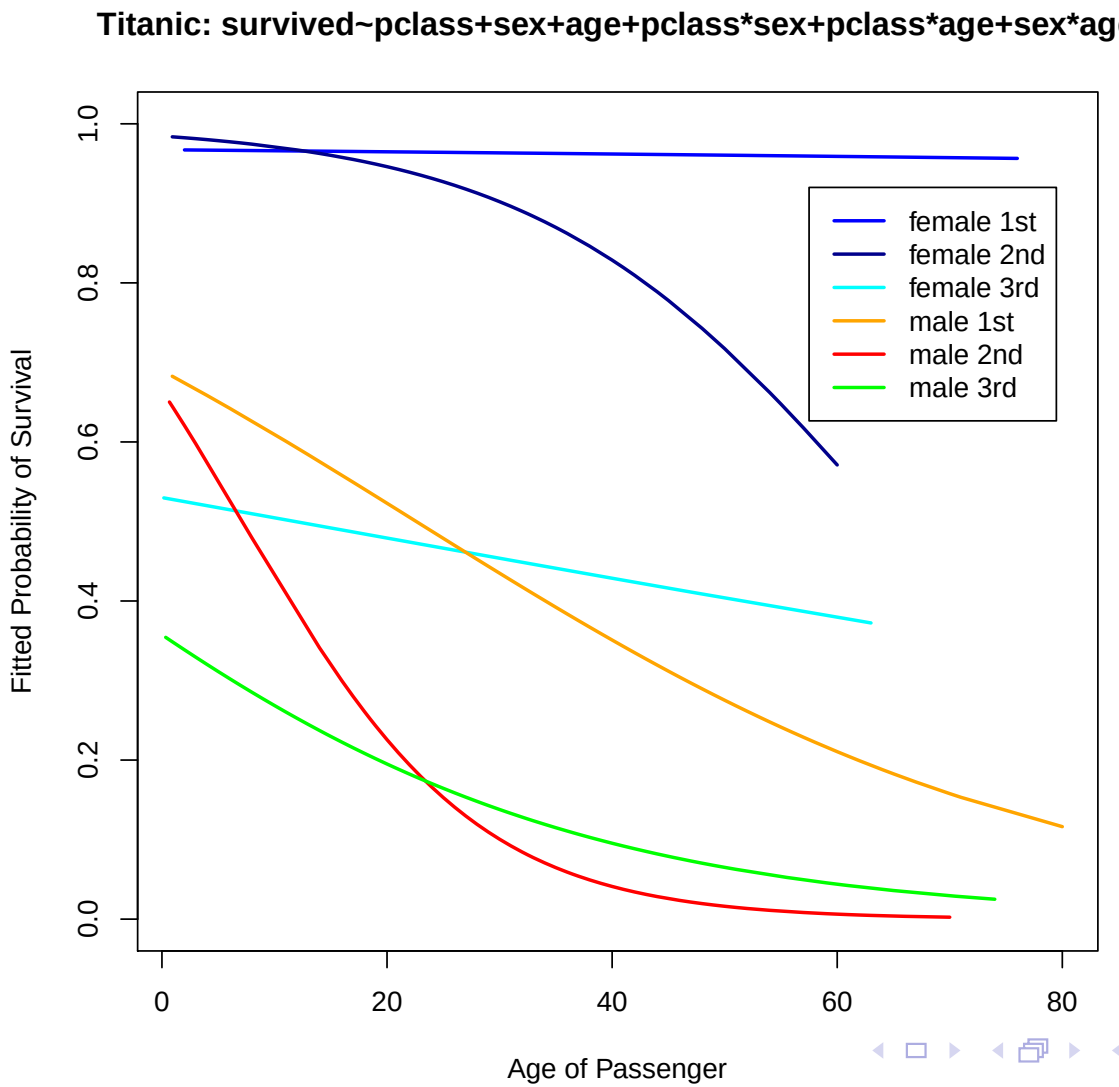


Parameter Estimates

Analysis of Maximum Likelihood Estimates

Effect/Parameter			df	Estimate	s.e.	Pr . Wald	ChiSq
Intercept			1	1.4269	.2773	26.4749	$<$ .01
pclass	1		1	0.6673	.4104	2.6433	.10
plcass	2		1	0.9925	.4061	5.9740	.01
sex	female		1	1.1283	.2559	19.4478	$<$ .01
age			1	-.0419	.0080	27.5016	$<$ .01
pclass*sex	1	female	1	0.1678	.1940	0.7480	.39
pclass*sex	2	female	1	0.6072	.1805	11.3190	$<$ .01
age*pclass	1		1	0.0223	.0108	4.2606	.04
age*pclass	2		1	-.0383	.0127	9.1143	$<$ .01
age*sex	female		1	0.0157	.0076	4.3075	.04

For Interpretation



Final Remarks. . . for now

Sample Size & Power.

In the text there are formulas of estimating the needed sample size to detect effects for a given significance level, power, and the effect size for the following cases:

- 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, effect size, etc.
- Should be use at the design stage of research

Exact Inference

- Maximum likelihood estimation of parameters works the best and statistical 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 the 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 depend only on the parameters that your're interested in estimating.
- 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.

Example Exact Inference

Since it is good for small samples, we will use ESR data ($n = 23$ and $y = 1$ 6 times).

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Std Error	Wald χ^2	Pr > ChiSq
Intercept	1	-12.7920	5.7964	4.8704	.0273
fibrin	1	1.9104	0.9710	3.8708	.0491
globulin	1	0.1558	0.1195	1.6982	.1925

Exact Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		Two-sided p-Value
fibrin	1.7274	0.9237	0.1648	3.8934	.0271
globulin	0.1054	0.1042	-0.0946	0.3644	.3396

Exact Inference

- Good for small data sets and relatively simple models; otherwise, it could take a very long time.
- How to do this in SAS:

```
title 'ESR Data:  exact';
proc logistic data=esr descending;
   model response=fibrin globulin;
   exact fibrin globulin / estimate=both;
run;
```

R Exact Inference

There was an R package, "elrm", that could do this but it was .....

```
Package elrm was removed from the CRAN repository.

Formerly available versions can be obtained from the
archive.

Archived on 2018-06-17 as check problems were not
corrected despite reminders.
```

If you want it, try this (didn't work for me):

```
require(devtools)
install_version("elrm", repos = "http://cran.r-project.org")
```

SAS for Logistic Regression

When the data are in a Subject $\times$ Variable matrix (i.e., 1 line per subject/case)

- PROC GENMOD: You need a variable for the number of cases (e.g., "ncases") that equals 1 for each individual.

$$\text{model } y/\text{ncases} = x1 \quad x2 \quad / \quad \text{link} = \text{logit dist} = \text{binomial};$$

- PROC LOGISTIC: You do not need the number of cases,

$$\text{model } y = x1 \quad x2 \quad / < \text{options desired} >;$$

Including Interactions

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 all lower order effects when you use $*$.

All useful: *model* $y/n = x1|x2|x3|x4 \quad @2$

This gives you all marginal effects and 2-way interactions, and @3 gives all marginal effects, 2–, and 3-way, etc.

Another R package: rms

The "Regression Modeling Strategies"

- I used this to get pseudo-R^2 in the Tianic example.
- It gives a lot more indices for evaluation (e.g., concordance index)
- Example use:
  model5.alt $\leftarrow$ lrm(survived $\sim$ pclass + sex + age,data=t)
- Run and check output.
- We will probability use this for ordinal logistic regression (I think)

Last 2 Comments on Logistic Regression

(. . . for now) Degrees of Freedom

$$\begin{aligned} df &= \text{num of logits} - \text{num of unique parameters} \\ &= \text{num of logits} - (\#\text{parameters} - \#\text{contraints}) \end{aligned}$$

High School and Beyond with school type and SES as nominal.

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

So

$$\begin{aligned} df &= (\#\text{school types}) \times (\#\text{SES levels}) \\ &\quad -(\#\text{unique parameters}) \\ &= (2 \times 3) - (1 + (2-1) + (3-1) = 2 \end{aligned}$$

For similar simple models: $df = (I-1)(J-1)$
Sometimes with numerical explanatory variables, you may want to first standardize them.