

Logistic Regression for Dichotomous Response Variables

Edps/Psych/Soc 589

Carolyn J. Anderson

Department of Educational Psychology



©Board of Trustees, University of Illinois

I Outline

In this set of notes:

- Review and Some Uses & Examples.
- Interpreting logistic regression models.
- Inference for logistic regression.
- Model checking.
- The Tale of the Titanic

Next set of notes will cover:

- Logit models for qualitative explanatory variables.
- Multiple logistic regression.
- Sample size & power.

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

I Additional References & Data

- Collett, D. (1991). *Analysis of Binary Data*.
- Hosmer, D.W., & Lemeshow, S. (1989). *Applied Logistic Regression*.
- McCullagh, P., & Nelder, J.A., (1989). *Generalized Linear Models*, 2nd Edition.
- SAS Institute (1995). *Logistic Regression Examples Using the SAS System*, Version 6.

Example data sets from SAS book are available via

- Anonymous ftp to <ftp.sas.com>.
- World wide web — <http://www.sas.com>

I Review of Logistic Regression

The logistic regression model is a generalized linear model with

- Random component: The response variable is **binary**. $Y_i = 1$ or 0 (an event occurs or it doesn't).

We are interesting in probability that $Y_i = 1$, $\pi(x_i)$.

The distribution of Y_i is **Binomial**.

- **Systematic component:** A linear predictor such as

$$\alpha + \beta_1 x_{1i} + \dots + \beta_j x_{ji}$$

The explanatory or predictor variables may be quantitative (continuous), qualitative (discrete), or both (mixed).

- Link Function: The log of the odds that an event occurs, otherwise known as the **logit**:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

Putting this all together, the logistic regression model is

$$\text{logit}(\pi(x_i)) = \log\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \alpha + \beta_1 x_{1i} + \dots + \beta_j x_{ji}$$

I Some Uses of Logistic Regression

To **model the probabilities** of certain conditions or states as a function of some explanatory variables.

To identify "Risk" factors for certain conditions (e.g., divorce, well adjusted, disease, etc.).

Diabetes Example (I got these data from *SAS Logistic Regression Examples* who got it from Friendly (1991) who got it from Reaven & Miller, 1979).

I (1) Example: Risk Factors

In a study of the relationship between various blood chemistry measures and diabetic status, data were collected from 145 nonobese adults who were diagnosed as Subclinical diabetic, Overt diabetic, or Normal

The possible explanatory variables:

- Relative weight (person's weight/expected weight given height or BMI).
- Fasting plasma glucose.
- Test plasma glucose intolerance.
- Plasma insulin during test (measure of insulin response to oral glucose).
- Steady state glucose (measure of insulin resistance).

I (2) Descriptive Discriminate Analysis

To describe differences between individuals from separate groups as a function of some explanatory variables —
descriptive discriminate analysis.

High School and Beyond data: The response variable is whether a student attended an academic program or a non-academic program (i.e., general or vocational/technical).

Possible explanatory variables include

- Achievement test scores (“continuous”) — reading, writing, math, science, and/or civics.
- Desired occupation (discrete–nominal) — 17 of them.
- Socio-Economic status (discrete–ordinal) — low, middle, high.

Goal/Purpose: Describe differences between those who attended academic versus non-academic programs.

I (4) Predict Probabilities

To **predict probabilities** that individuals fall into one of 2 categories on a dichotomous response variable as a function of some set of explanatory variables. This covers lots of studies (from epidemiological to educational measurement).

Example: ESR Data from Collett (1991).

A healthy individual should have an erythrocyte sedimentation rate (ESR) less than 20 mm/hour. The value of ESR isn't that important, so the response variable is just

$$Y_i = \begin{cases} 1 & \text{if ESR} < 20 & \text{or healthy} \\ 0 & \text{if ESR} \geq 20 & \text{or unhealthy} \end{cases}$$

The possible explanatory variables were

- Level of plasma fibrinogen (gm/liter).
- Level of gamma-globulin (gm/liter).

I (4) Predict Probabilities

An example from Anderson, Kim & Keller (2013): PIRLS data from US

Response variable: Response of student to a question about how often they look up information on the computer for school
 (“Every day or almost every day”, “Once or twice a week”, “Once or twice a month”, “Never or almost never”)

Explanatory variables: gender, how much time they spend per day reading for homework, screen time per day, availability of computers in their school, location of school, percent of students at school that get free or reduced price lunch, school climate).

Complications: multilevel structure, design/sampling weights, and missing data.

I (5) Classify Individuals

To classify individuals into one of 2 categories on the basis of the explanatory variables.

Effron (1975), Press & Wilson (1978), and Amemiya & Powell (1980) compared logistic regression to discriminant analysis (which assumes the explanatory variables are multivariate normal at each level of the response variable).

Eshan Bokhari (2014): Compared logistic regressions & discriminant analysis for identifying who will commit violent act. (Bokari & Hubert method seems to be best).

I (6) Discrete Choice

To analyze responses from discrete choice studies (estimate choice probabilities).

From *SAS Logistic Regression Examples* (hypothetical).

Chocolate Candy: 10 subjects presented 8 different chocolates choose which one of the 8 is the one that they like the best. The 8 chocolates consisted of 2^3 combinations of

- Type of chocolate (milk or dark).
- Center (hard or soft).
- Whether it had nuts or not.

The response is which chocolate most preferred.

I (6) Discrete Choice (continued)

The different names for this particular logit model are

- The multinomial logit model.
- McFadden's model.
- Conditional logit model.

This model is related to Bradley-Terry-Luce choice model.

This model is used

- To analyze choice data and use characteristics of the objects or attributes of the subject as predictors of choice behavior.
- In marketing research to predict consumer behavior.
- As an alternative to conjoint analysis.

I (7) Social Network Analysis

- Data often consist of individuals (people, organizations, countries, etc.) within a group or network upon which relations are recorded (e.g., is friends with, talks to, does business with, trades, etc).
- The relations can be
 - Undirected (e.g., is biologically related to)
 - Directed (e.g., asks advise from, gives money to)
- Example: Data from Parker & Asher (1993). Children in 35 different classes were asked who they were friends with (in the class). Other measures were also taken, including gender, race, a loneliness or “connectedness” measure, and others.
- This sort of data is often organized in a “sociomatrix”, which consists of a matrix of binary random variables:

$$X_{ij} = \begin{cases} 1 & \text{if } i \text{ chooses } j \\ 0 & \text{otherwise} \end{cases}$$

I (7) Social Network Analysis

- Problem: A family of models used to analyze such data is (Poisson) log-linear (Wasserman & Faust 1995); however, these models make the assumption that the pairs of individuals (“dyads”) are independent, which has been the major criticism of these models.
- Solution: Logit/Logistic regression where you model the odds of the existence of a relation between two actors conditioning upon the present/absence of ties between actors in the rest of the network (see Wasserman & Pattison, 1996; Anderson, Wasserman & Crouch, 1999) Currently favored estimation methods for “p*” or “exponential random graphical models” (EGRMS) are Bayesian.

I (8) Pseudo-likelihood Estimation & Others

- Log-linear type models have estimation problems for really large problems.
e.g. My research on log-multiplicative association models as latent variable models (graphical measurement models).
- Can use maximize pseudo-likelihood by maximum likelihood estimation of an appropriate logistic regression model, which is a discrete choice model.

(9) There are others! e.g., DIF, estimate standard IRT models, survival/event history analysis

Dated but still good: David Strauss. (1992). The many faces of logistic regression *American Statistician*. . . more uses since Strauss's paper.

I Interpreting logistic regression models

The model

$$\text{logit}(\pi(x)) = \text{logit}\pi(x) = \alpha + \beta x$$

or alternatively in terms of $\pi(x)$

$$\pi(x) = \frac{\exp\{\alpha + \beta x\}}{1 + \exp\{\alpha + \beta x\}}$$

In considering the various interpretations of logistic regression, we'll use the High School and Beyond data (for now).

- Response: $Y_i = 1$ if the student attended an academic program, and 0 if the student attended a non-academic program.
- Explanatory: x_i = student's mean of five achievement test scores: reading, writing, math, science, and civics.

Each test on a T -score scale (i.e., mean = 50, and standard deviation = 10).

I HSB Example

HSB Example: The simplest model, the **linear probability** model (i.e., link=identity, and distribution is Binomial).

$$\hat{\pi}(x_i) = -.9386 + .0281x_i$$

Problems:

- We get some negative fitted values and some greater than 1. (R glm completely fails)
- The rate of change in the probability of attending an academic program is not constant across possible values of the mean Achievement T -scores.

Logistic regression (i.e., logit link and Binomial distribution).

The estimated model

$$\begin{aligned}\text{logit}(\hat{\pi}_i) = \text{logit}\hat{\pi}_i &= \hat{\alpha} + \hat{\beta}x_i \\ &= -7.0548 + .1369x_i\end{aligned}$$

Note: ASE for $\hat{\alpha}$ equals .6948 and ASE for $\hat{\beta}$ equals .0133.

I HSB: Observed and Fitted

To help “see” how well this model does, the mean scores were grouped into 11 categories (a lot fewer than the 531 unique math scores).

Group	# attend acad	# cases	Observed proportion	Predicted Prob Sum	Predicted Prob Equation	Pred #acad
$x < 40$	8	46	.17	.12	.13	5.64
$40 \geq x < 45$	18	87	.20	.19	.23	16.75
$45 \geq x < 47$	14	44	.31	.27	.32	12.82
$47 \geq x < 49$	17	43	.39	.36	.38	15.46
$49 \geq x < 51$	18	50	.36	.40	.45	20.00
$51 \geq x < 53$	22	50	.44	.49	.52	24.98
$53 \geq x < 55$	34	58	.58	.49	.58	28.52
$55 \geq x < 57$	23	44	.52	.56	.65	24.70
$57 \geq x < 60$	35	68	.51	.58	.72	39.52
$60 \geq x < 65$	56	78	.71	.75	.82	58.62
$65 \geq x$	26	32	.81	.80	.89	25.68

I Where...

- “Predicted # academic” = $\sum_i \hat{\pi}(x_i)$, where the sum is over those values of i that correspond to observations with x_i within each of math score categories.
- “Predicted probability— Sum” = $\sum_i \hat{\pi}(x_i)/(\# \text{ cases})$.
- “Predicted probability— Equation”

$$= \exp(-7.0548 + .1369(\overline{\text{achieve}_i})) / (1 + \exp(-7.0548 + .1369(\overline{\text{achieve}_i})))$$

I SAS Code to Group

This will give approximately equal numbers per group

```
/* Produce quantiles for achievement scores */
proc rank data=preds groups=10
    out=group(keep=grp achieve academic);
    var achieve;
    ranks grp;
run;

title 'To show that (nearly) equal groups were created ';
proc freq data=group;
    tables grp;
```

I Show on Equal Grouping

Rank for Variable Grp

grp	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	60	10.00	60	10.00
1	60	10.00	120	20.00
2	60	10.00	180	30.00
3	60	10.00	240	40.00
4	60	10.00	300	50.00
5	60	10.00	360	60.00
6	60	10.00	420	70.00
7	59	9.83	479	79.83
8	61	10.17	540	90.00
9	60	10.00	600	100.00

I File Created by PROC RANK

achieve	academic	grp
39.42	0	0
41.32	0	1
46.56	0	2
39.00	0	0
42.40	0	1
43.86	0	1
58.84	1	7
47.60	1	3
47.20	1	3
44.76	1	2
49.32	1	3
45.02	0	2
67.94	1	9

I After you Have GRP Variable

- Sort the data by grp:

```
proc sort data=group1;
  by grp;
```

- Compute the sums:

```
proc means data=group1 sum;
  by grp;
  var academic count fitted;
  output out=grpfit sum=num__aca num__cases fit2;
```

- One more data step:

```
data done;
  set grpfit;
  p=num__aca/num__cases;
  pp = fit2/num__cases;
run;
```

I Grouping using R

- I assume that you have fitted logit model:

```
logit.mod <- glm(program ~ achieve, data=hsb,
                  family=binomial(link="logit"))
```

- Decide on levels/cutpoints:

```
levels <- c(-Inf, 40, 45, 47, 49, 51, 53, 55, 57, 60,
            65, Inf)
```

- Give names/labels to groupings

```
labels <- c("<40", "40-45", "45-47", "47-49", "47-51",
            "51-53", "53-55", "55-57", "57-60", "60-65", ">65")
```

- Create a new variable in hsb with group variable

```
hsb$grp.achieve = cut(hsb$achieve, levels,
                      labels = labels)
```

- If you want to check and will want this

```
grp.n <- tables(hsb$grp.achieve)
```

I Grouping using R

- Number of non and academic students per group

```
data.table ← table(hsb$grp.achieve, hsb$program)
```

- Observed proportions

```
obs.p ← data.table[,2] / (data.table[,1]
                          + data.table[,2])
```

- Sum predicted probabilities

```
prob.sum ← aggregate(hsb$logit.fit,
                      by=list(hsb$grp.achieve), FUN=sum) / grp.n
```

- Using equation to get predicted probabilities using equation on grouped data

```
p.eq ← exp(-7.05477 + 0.13691*grp.mean[,2]) / (1 +
exp(-7.05477 + 0.13691*grp.mean[,2]))
```

I Finishing Up

Put into table like in lecture notes

- Predicted number (i.e., frequency)

```
pred.n ← p.eq*grp.n
```

```
pred.n ← matrix(pred.n,nrow=11,ncol=1)
```

- `summary.p19 ← cbind(grp.n,grp.mean[,2],
data.table[,2], obs.p,prob.sum[,2], p.eq,pred.n)`
- `summary.p19 ← as.data.frame(summary.p24)`
- `names(summary.p19) ← c("range", "cases ", "mean",
"academic", "obs p", "sum pi", "equation pi", "pred
freq")`
- `summary.p19`

I Interpreting β

Recall that β determines the rate of change of the curve of $\pi(x)$ (plotted with values of x along the horizontal axis) such that

If $\beta > 0$, then the curve increases with x

If $\beta < 0$, then the curve decreases with x

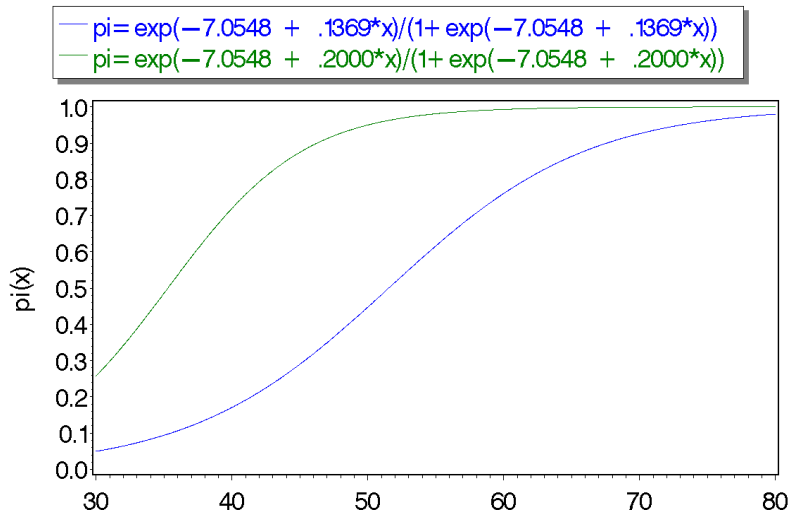
If $\beta = 0$, then curve is flat (horizontal)

To see how the curve changes as β changes:

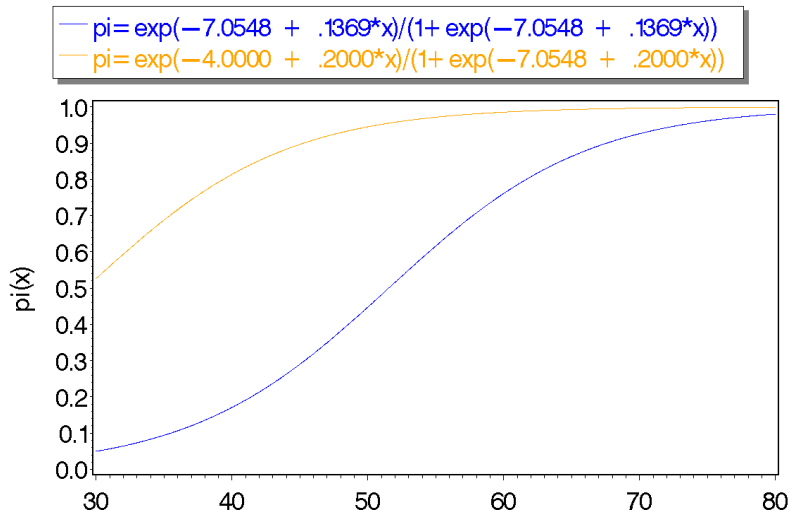
Curve on the left: $\text{logit}(\pi(x)) = -7.0548 + .2000x$

Curve on the right: $\text{logit}(\pi(x)) = -7.0548 + .1369x$

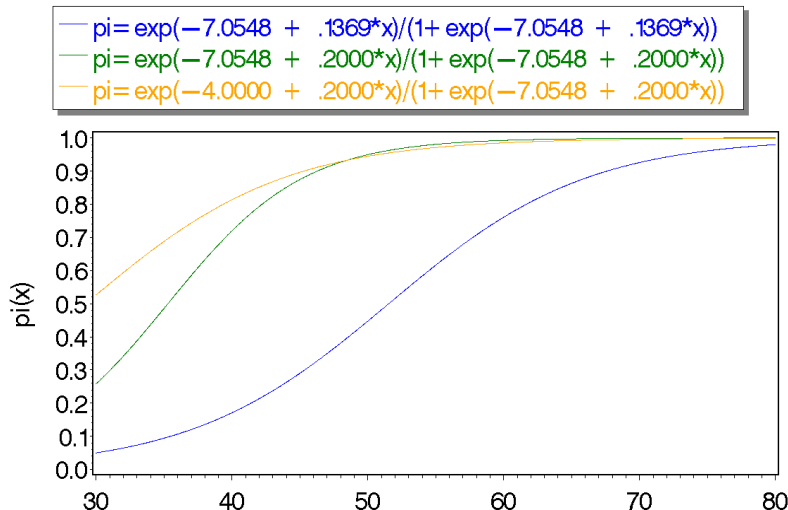
I Figure: Interpreting β



I Different α 's



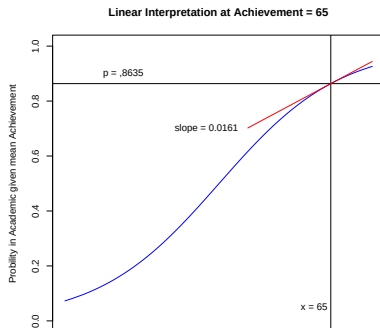
I Different α 's & Different β 's



I Linear Approximation Interpretation

To illustrate this, we'll use the model estimated for the High School and Beyond Data,

$$\hat{\pi}(x_i) = \frac{\exp\{-7.0548 + .1369x_i\}}{1 + \exp\{-7.0548 + .1369x_i\}}$$



I Linear Approximation Interpretation

- Since the function is curved, the change in $\pi(x)$ for a unit change in x is not constant but varies with x .
- At any given value of x , the rate of change corresponds to the slope of the curve — draw a line tangent to the curve at some value of x , and slope (rate of change) equals

$$\beta\pi(x)(1 - \pi(x))$$

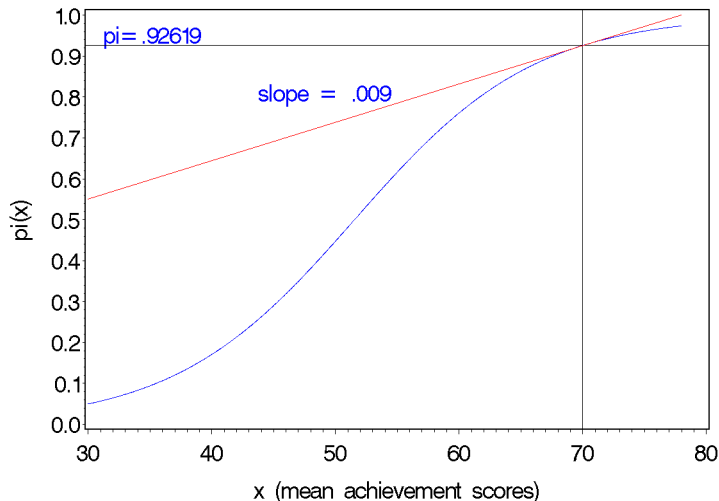
- For example, when the math achievement score x equals 65,

$$\hat{\pi}(65) = \frac{\exp\{-7.0548 + .1369(65)\}}{1 + \exp\{-7.0548 + .1369(65)\}} = .8634$$

$$1 - \hat{\pi}(65) = 1 - .8634 = .1366$$

and the slope equals $\beta\pi(65)(1 - \pi(65)) = (.1366)(.8634) = .0161$.

I Linear Approximation at $x=70$



I Linear Approximation Interpretation

- The slope is greatest when $\pi(x) = (1 - \pi(x)) = .5$; that is, when

$$x = -\alpha/\beta = -(-7.0548)/.1369 = 51.53$$

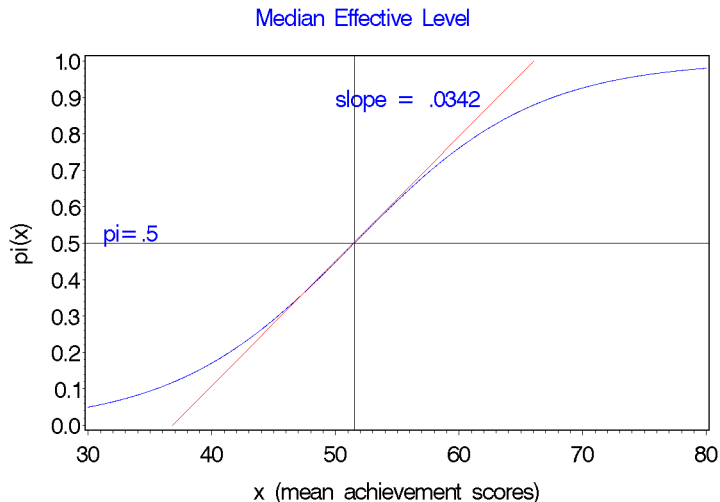
$$\hat{\pi}(51.53) = (1 - \hat{\pi}(51.53)) = .5$$

and slope at $x = 51.53$ is $(.1369)(.5)(.5) = .034$

- The value of $x = -\alpha/\beta$ is called the “**median effective level**” or EL_{50} (for short), because it is the point at which each event is equally likely.
- Some other values:

x_i	$\hat{\pi}_i$	$1 - \hat{\pi}_i$	Slope at x_i
70	.9261	.0739	.009
60	.7612	.2388	.025
52	.5160	.4840	.03419
51.5325	.5000	.5000	.03423
43.065	.2388	.7612	.025

I Linear Approximation at $x=51.53$



I Odds Ratio Interpretation

A somewhat simpler & more natural interpretation of logit/logistic regression models,

$$\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{(1 - \pi(x))}\right) = \alpha + \beta x$$

- Taking the exponential of both sides,

$$\frac{\pi(x)}{1 - \pi(x)} = \exp\{\alpha + \beta x\} = e^\alpha e^{\beta x}$$

- This is a model for odds; Odds change multiplicatively with x .
- A 1 unit increase in x leads to an increase in the odds of e^β . So the odds ratio for a 1 unit increase in x equals

$$\frac{\pi(x+1)/(1 - \pi(x+1))}{\pi(x)/(1 - \pi(x))} = \frac{e^\alpha e^{\beta x} e^\beta}{e^\alpha e^{\beta x}} = e^\beta$$

I Odds Ratio Interpretation (continued)

- When $\beta = 0$, $e^0 = 1$, so the odds do not change with x .
- The logarithm of the odds changes linearly with x ; however, the logarithm of odds is not an intuitively easy or natural scale to interpret.

I HSB: odds ratio interpretation

- Since $\hat{\beta} = .1369$, a 1 unit increase in mean achievement test scores leads to an odds ratio of

$$\text{odds ratio for } (\Delta x = 1) = e^{.1369} = 1.147$$

- The odds of having attended an academic program given a mean achievement score of $x + 1$ is 1.147 times the odds given a mean achievement score of x .
- If x changes by 10 (1 s.d. on the T -score scale), then the odds ratio is

$$\begin{aligned} \text{odds ratio for } (\Delta x = 10) &= \frac{e^{\alpha} e^{\beta(x+10)}}{e^{\alpha} e^{\beta(x)}} \\ &= \frac{e^{\alpha} e^{\beta x} e^{\beta(10)}}{e^{\alpha} e^{\beta(x)}} = e^{\beta(10)} \end{aligned}$$

- For our example, $e^{.1369(10)} = 3.931$
- Unlike the interpretation in terms of probabilities (where the rate of change in $\pi(x)$ is not constant for equal changes in x), the odds ratio interpretation leads to constant rate of change.

I Random Explanatory variable & Fixed Response

- This happens in retrospective studies (e.g., case–controls)
- From Hosmer & Lemeshow (1989): In a study investigating the risk factors for low birth weight babies, the risk factors considered
 - Race
 - Smoking status of mother
 - History of high blood pressure
 - History of premature labor
 - Presence of uterine irritability
 - Mother's pre-pregnancy weight
- The 56 women who gave birth to low weight babies in this study were matched on the basis of age with a randomly selected control mother (i.e. each control gave birth to a normal weight baby and was the same age as the “case” mother).

I Example Continued

- If the distribution of explanatory variables /risk factors is different for the case & control moms, then this is evidence of an association between low birth weight & the risk factors.
- The estimated coefficients of an explanatory variable can be used to **estimate the odds ratio**. Note: this only works for logit/logistic regression model for binary data, and does *not* work for linear & probit models for binary data.
- You'll have to wait for the results check reference or analyze the data yourself (the data are in the MASS package in R and the data set is named "birthwt").

I A Special Case

- Whether a logistic regression model is a good description of a data set is an empirical question, except for one particular case. . .
- The logistic regression model necessarily holds when
 - The distribution of X for all those with $Y = 1$ is $\mathcal{N}(\mu_1, \sigma^2)$.
 - The distribution of X for all those with $Y = 0$ is $\mathcal{N}(\mu_0, \sigma^2)$.
- Do these assumptions sound familiar?
- If these 2 conditions hold, then
 - $\pi(x)$ follows a logistic regression curve,

$$\text{logit}(\pi(x)) = \alpha + \beta_1 x$$

- The sign of β is the same as the sign of $\mu_1 - \mu_0$.
- If the variances are quite different, then a logistic regression model for $\pi(x)$ that also contains a quadratic term is likely to fit the data well.

$$\text{logit}(\pi(x)) = \alpha + \beta_1 x_1 + \beta_2 x_1^2$$

I Inference for logistic regression

Or the significance and size of effects

- 1 Confidence intervals for parameters.
- 2 Hypothesis testing.
- 3 Distribution of probability estimates.

(1) and (2) will follow much like what we did for Poisson regression. (3) will be a bit different.

I Confidence Intervals in Logistic Regression

- Since we use maximum likelihood estimation, for large samples, the distribution of parameter estimates is approximately normal.
- A large sample $(1 - \alpha)100\%$ confidence interval for β is

$$\hat{\beta} \pm z_{\alpha/2}(ASE)$$

where α here refers to the significance level (and not the intercept of the model).

- Example (High School and Beyond): A 95% confidence interval for β , the coefficient for mean achievement test scores is

$$.1369 \pm 1.96(.0133) \longrightarrow (.1109, .1629)$$

- To get an interval for the effect of mean achievement score on the odds, that is for e^{β} , we simply take the exponential of the confidence interval for β .
- $(e^{.1109}, e^{.1629}) \longrightarrow (1.1173, 1.1770)$

I Confidence Intervals for linear approximation

i.e., for $\beta\pi(x)(1 - \pi(x))$,

- Multiply the endpoints of the interval for β by $\pi(x)(1 - \pi(x))$.
- For $\pi(x) = .5$, so $\pi(x)(1 - \pi(x)) = .25$, a 95% confidence interval for $\beta\pi(x)(1 - \pi(x))$, the slope when $X = x$, is

$$(.25)(.1109), (.25)(.1629) \longrightarrow (.0277, .0407)$$

- So the incremental rate of change of $\pi(x)$ when $x = -\hat{\alpha}/\hat{\beta} = 51.5325$ is an increase in probability of .0277 to .0407.

I Hypothesis Testing: $H_O : \beta = 0$

i.e., X is not related to response.

- ① Wald test
- ② Likelihood ratio test

Wald test: For large samples,

$$z = \frac{\hat{\beta}}{ASE}$$

is approximated $\mathcal{N}(0, 1)$ when $H_O : \beta = 0$ is true — **R reports this**.
So for 1-tailed tests, just refer to standard normal distribution.

$$\text{Wald statistic} = \left(\frac{\hat{\beta}}{ASE} \right)^2$$

which if the null is true is approximately chi-square distributed with $df = 1$. — **SAS reports this**.

I HSB: Wald Test

$H_O : \beta = 0$ (i.e., mean achievement is not related to whether a student attended an academic or nonacademic program)

versus

$H_A : \beta \neq 0$.

$$\text{Wald statistic} = \left(\frac{.1369}{.0133} \right)^2 = (10.29)^2 = 106$$

which with $df = 1$ has a very small p -value.

I Likelihood ratio test statistic

...the more powerful alternative to Wald test.

$$\text{test statistic} = \text{LR} = -2(L_O - L_1)$$

where L_O is the log of the maximum likelihood for the model

$$\text{logit}(\pi(x)) = \alpha$$

and L_1 is the log of the maximum likelihood for the model

$$\text{logit}(\pi(x)) = \alpha + \beta x$$

If the null is true, then the likelihood ratio test statistic is approximately chi-square distributed with $df = 1$.

HSB Example:

$$\begin{aligned} \text{LR} &= -2(L_O - L_1) \\ &= -2(-415.6749 - (-346.1340)) = 139.08, \quad df = 1, \quad p < .01 \end{aligned}$$

I SAS: Wald & Likelihood Ratio

The easier way to get LR test statistic in SAS: “type3” as a model option:

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
achieve	1	139.08	< .0001

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-7.05	0.69	103.10	< .0001
achieve	1	0.14	0.01	106.41	< .0001

Why is the Wald statistic “only” 106.41, while the likelihood ratio statistic is 139.08 and both have the same *df* & testing the same hypothesis?

I R: Wald & Likelihood Ratio

In R, table of parameter estimates:

Coefficients:

	Estimate	Std. Error	z value	Pr(< z)
(Intercept)	-7.05477	0.69475	-10.15	<2e-16 ***
achieve	0.13691	0.01327	10.32	<2e-16 ***

Easy way to get Likelihood ratio, `anova(logit.mod)`

	Df	Deviance	Resid.	Df	Resid. Dev
NULL				599	831.35
achieve	1	139.08		598	692.27

(i.e., take difference between Null and Residual deviance).

for p -value, use `1-pchisq(139.08,1)`

or `1-pchisq(logit.mod$deviance,1)`

I Confidence Intervals for Probability Estimates

Our estimated probability for $X = x$,

$$\hat{\pi}(x) = \frac{\exp\{\hat{\alpha} + \hat{\beta}x\}}{1 + \exp\{\hat{\alpha} + \hat{\beta}x\}}$$

Want confidence intervals for $\hat{\pi}(x)$ using the estimated model.

HSB example with Mean achievement score as the explanatory variable:
Suppose we're interested in the probability when achievement score
= 51.1. the estimated probability (or propensity) that a student attended
an academic program equals

$$\begin{aligned}\hat{\pi}(51.1) &= \frac{\exp\{-7.0548 + .1369(51.1)\}}{1 + \exp\{-7.0548 + .1369(51.1)\}} \\ &= e^{-.05842} / (1 + e^{-.05842}) = .4854\end{aligned}$$

I CI for Probability Estimates

From PROC GENMOD, a 95% confidence interval for the true probability when $x = 51.1$ is

(.440, .531)

If you use SAS/GENMOD with the “obstats” option, the table created by obstats contains:

Column Label	Translation	HSB Example (for $x = 51.1$)
Pred	$\hat{\pi}(x)$	0.4853992
Xbeta	$\text{logit}(\hat{\pi}(x)) = \hat{\alpha} + \hat{\beta}x$	-0.05842
Std	$\sqrt{\widehat{\text{Var}}(\text{logit}(\hat{\pi}(x)))}$	0.0927311
Lower	Lower value of 95% CI for $\pi(x)$	.4402446
Upper	Upper value of 95% CI for $\pi(x)$	.5307934

and how SAS got Std, Lower, and Upper...

I Computational Details of CI for $\hat{\pi}(x_i)$

Find a confidence interval for the $\text{logit}(\pi(x))$ and then transform is back to probabilities.

To compute a confidence interval for the logit, $\alpha + \beta x$, we need an estimate of the variance of $\widehat{\text{logit}}(\pi(x))$, that is,

$$\widehat{\text{var}}(\hat{\alpha} + \hat{\beta}x)$$

which is equal to

$$\widehat{\text{var}}(\hat{\alpha} + \hat{\beta}x) = \widehat{\text{var}}(\hat{\alpha}) + x^2\widehat{\text{var}}(\hat{\beta}) + 2x\widehat{\text{cov}}(\hat{\alpha}, \hat{\beta})$$

The estimated variances and covariances are a by-product of the estimation procedure that SAS uses. The **CovB** option in the **model** statement requests that the estimated variance/covariance matrix of estimates parameter be printed (in the listing file or output window).

I Computational Details of CI for $\hat{\pi}(x_i)$

Estimated Covariance Matrix in SAS output from the **covb** option to the **MODEL** statement:

	Prm1	Prm2
Prm1	0.48271	-0.009140
Prm2	-0.009140	0.0001762

Estimated Covariance Matrix of Estimated Parameters:

	$\hat{\alpha}$	$\hat{\beta}$
$\hat{\alpha}$	0.48271	-0.009140
$\hat{\beta}$	-0.009140	0.0001762

Note: ASE of $\hat{\beta} = \sqrt{.0001762} = .0133$, as previously given.

I Computing CI for $\hat{\pi}(x_i)$

So for $x = 51.1$,

$$\begin{aligned}\widehat{\text{var}}(\hat{\alpha} + \hat{\beta}x) &= \widehat{\text{var}}(\hat{\alpha}) + x^2\widehat{\text{var}}(\hat{\beta}) + 2x\widehat{\text{cov}}(\hat{\alpha}, \hat{\beta}) \\ &= .48271 + (51.1)^2(.0001762) + 2(51.1)(-.009140) = .008697\end{aligned}$$

$$\text{and } \sqrt{\widehat{\text{Var}}(\hat{\alpha} + \hat{\beta}x)} = \sqrt{.008697} = .0933$$

A 95% confidence interval for the true **logit** when $x = 51.1$ is

$$\begin{aligned}\hat{\alpha} + \hat{\beta}x &\pm 1.96\sqrt{\widehat{\text{Var}}(\hat{\alpha} + \hat{\beta}x)} \\ -0.0584 &\pm 1.96(.0933) \rightarrow (-.2413, .1245)\end{aligned}$$

and finally to get the 95% confidence interval for the true **probability** when $x = 51.1$, transform the endpoints of the interval for the logit to probabilities:

$$\left(\frac{\exp(-.2413)}{1 + \exp(-.2413)}, \frac{\exp(.1245)}{1 + \exp(.1245)} \right) \rightarrow (.44, .53)$$

I R: Computing CI for $\hat{\pi}(x_i)$

```
# Find min and max of predictor/explanatory variable
summary(hsb$sachieve)

# Create object that has values from min to max
achievement.range ← seq(from=32,to=350,by=1)

# This gets the fitted probability and it's se
logit.modt ← predict(logit.mod,
                     whatever ← data.frame(achieve = achievement.range),
                     type="response", se.fit=TRUE )

# To see what logit.modt has in it
names(logit.modt)

# Finally create interval
upper ← logit.modt$fit + 1.96*logit.modt$se.fit
lower ← logit.modt$fit - 1.96*logit.modt$se.fit
```

I R: Computing CI for $\hat{\pi}(x_i)$

Alternative is to use

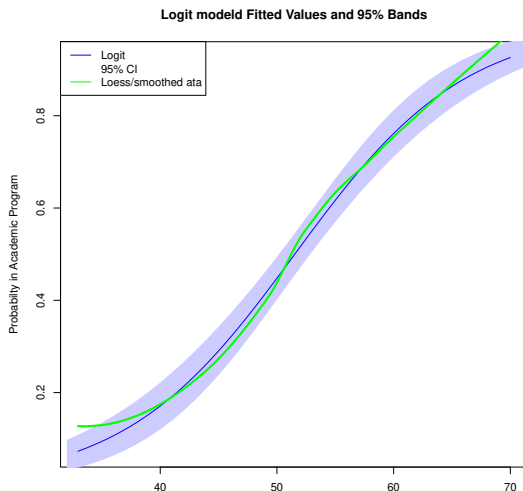
```
vcov(logit.mod)
```

which yields,

	(Intercept)	achieve
(Intercept)	0.482683791	-0.0091394759
achieve	-0.009139476	0.0001761518

And use `logit.mod$coefficients`

I CI bands for Fitted



- SLIDE -

I Model vs Non-model based CI for π

- Model based CI's are “better” than the non-model based ones.
- e.g., mean achievement 58.84, non-model based
 $n = 2$, $\# \text{ academic} = 1$, and $p = 1/2 = .5$
- Whereas the model based estimate equals $\hat{\pi}(58.84) = .73$.
- Can't even compute a 95% confidence interval for π using the (non-model) based sample proportion?
- With the logistic regression model, the model based interval is (.678, .779).
- The model confidence interval will tend to be much narrower than ones based on the sample proportion p , because... e.g., the estimated standard error of p is

$$\sqrt{p(1-p)/n} = \sqrt{.5(.5)/2} = .354$$

while the estimated standard error of $\hat{\pi}(58.84)$ is .131.

- The model uses all 600 observations, while the sample proportion only uses 2 out of the 600 observations.

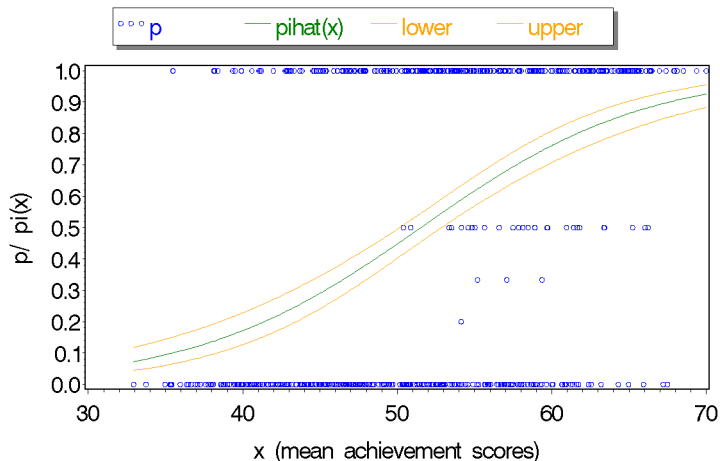
I Comments regarding of Model based Estimates

- Models do not represent the *exact* relationship between $\pi(x)$ and x .
- As the sample size increases (i.e., n_i for each x_i), $\hat{\pi}(x_i)$ does not converge to the true probabilities; however, p does.
- The extent to which the model is a good approximation of the true probabilities, the model based estimates are closer to the true values than p and the model based have lower variance.
- Models “smooth” the data.
- Observed proportions p versus math scores and model estimates of $\pi(x)$ versus math scores (next slide).
- For the HSB example, most of the sample p_i 's have n_i 's of 0 or 1 (largest is 5).

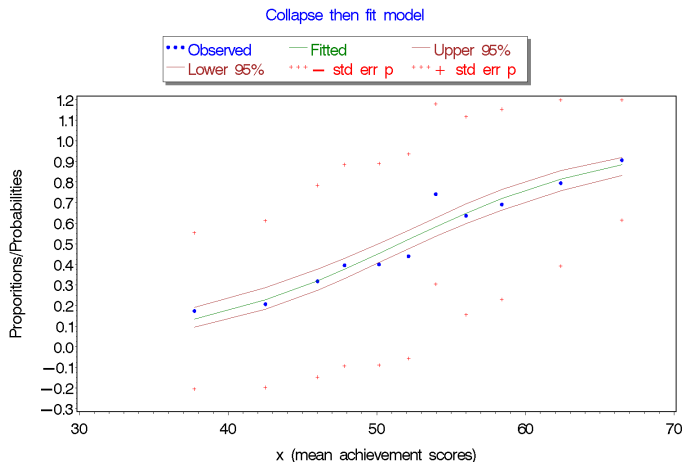
The above results provide additional incentive to investigate whether our model is a good one; that is, does the model approximate the true relationship between $\pi(x)$ and x .

I Model based vs Non

Model fit then collapsed



I Model based vs Non



I Model checking

Outline:

- ① Goodness-of-fit tests for continuous x .
 - ① Group observed counts & fitted counts from estimated model.
 - ② Group observed counts and then re-fit the model.
 - ③ Hosmer & Lemeshow goodness-of-fit test.
- ② Likelihood ratio model comparison tests.
- ③ Residuals.
- ④ Measures of influence.
- ⑤ ROC

I Goodness-of-fit tests when x “continuous”

- If most of the estimated counts are ≥ 5 , then G^2 and X^2 are approximately chi-squared distributed with df = “residual df ” where

$$\text{residual } df = \# \text{ sample logits} - \# \text{ model parameters}$$
- If the p -value is small, then we have evidence that the model does **not** fit the data.
- **However**, “continuous” explanatory variable creates a problem (i.e., X^2 and G^2 fail to have approximate χ^2 sampling distributions).
- HSB example:
 - There are 531 different values for achievement.
 - For each achievement value, we have $y_i \leq 5$.
 - If we considered the data in the form of a (531×2) contingency table of achievement $\times$ program type, many of the $(531)(2) = 1062$ cells of this table would be empty and contain very small cell counts.
 - There are only 600 observations/students.

I Large Sample Theory Requirements

Large sample theory for the goodness of model fit tests is violated in two ways:

- ① Most of the fitted cell counts are very small.
- ② As the number of students increase, the number of possible scores would also increase, which means that sample size effects the number of cells in the table.

So, what did we do with Poisson regression?...

I Fit Model then Group

First the model was fit to the data and then the observed and fitted counts were grouped. (grouped to approximately equal cases per category or group equally spaced along x).

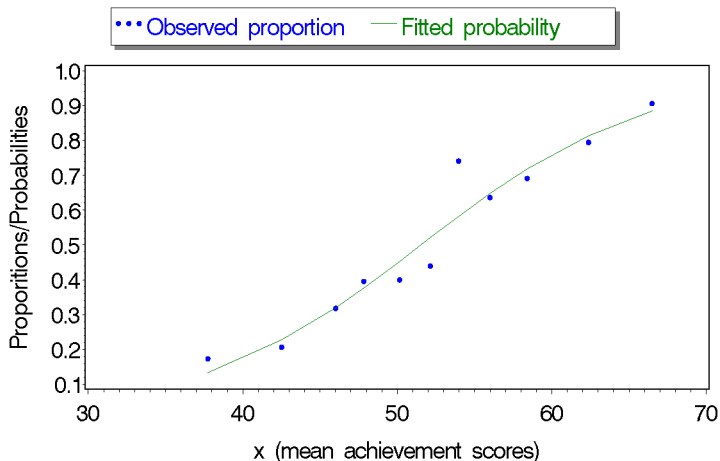
Group	Mean score	Observed Data			Fitted Values	
		# attend acad	# attend non-acad	# cases in group	# attend acad	# attend non-acad
$x < 40$	37.77	8	38	46	6.48	39.52
$40 \leq x < 45$	42.60	18	69	87	16.86	70.14
$45 \leq x < 47$	46.03	14	30	44	11.98	32.02
$47 \leq x < 49$	47.85	17	26	43	13.97	29.03
$49 \leq x < 51$	50.12	18	32	50	17.41	32.59
$51 \leq x < 53$	52.12	22	28	50	21.44	28.56
$53 \leq x < 55$	53.95	34	24	58	24.21	33.79
$55 \leq x < 57$	56.05	23	21	44	20.86	23.15
$57 \leq x < 60$	58.39	35	33	68	33.45	34.55
$60 \leq x < 65$	62.41	56	22	78	50.52	27.48
$65 \leq x$	66.56	26	6	32	22.73	9.27

I Plot of Counts



I Plot of Proportions & Probabilities

Model fit then collapse



I Fitting Model and then Collapse

Test statistics for goodness of fit:

$$X^2 = \sum_{\text{groups}} \sum_{\text{program}} \frac{(\text{observed} - \text{fitted})^2}{\text{fitted}}$$

$$G^2 = 2 \sum_{\text{groups}} \sum_{\text{program}} \text{observed} \log(\text{observed}/\text{fitted})$$

and

$$df = \# \text{ group} - \# \text{ parameters}$$

Using the values, we get

Statistic	df	Value	" p -value"
X^2	$(11 - 2) = 9$	9.40	
G^2	$(11 - 2) = 9$	9.72	

I Doing this in SAS

Step 1: Fit the model to the data:

```
PROC GENMOD data=hsb;
  model academic/ncases = achieve / link=logit
    dist=binomial obstats type3 covb;
  output out=preds pred =fitted;
```

Step 2: Use PROC FREQ to decide on cut-points:

```
PROC FREQ data=preds;
  tables achieve / nopercnt norow nocol list;
```

I Doing this in SAS

Step 3: Use cut-points to make grouping variable:

```
DATA group1;
```

```
  set preds;
```

```
  if achieve<40 then grp=1;
```

```
  else if achieve>=40 and achieve<45 then grp=2;
```

```
  else if achieve>=45 and achieve<47 then grp=3;
```

```
  else if achieve>=47 and achieve<49 then grp=4;
```

```
  else if achieve>=49 and achieve<51 then grp=5;
```

```
  else if achieve>=51 and achieve<53 then grp=6;
```

```
  else if achieve>=53 and achieve<55 then grp=7;
```

```
  else if achieve>=55 and achieve<57 then grp=8;
```

```
  else if achieve>=57 and achieve<60 then grp=9;
```

```
  else if achieve>=60 and achieve<65 then grp=10;
```

```
  else if achieve>=65 then grp=11;
```

I For equal groups sizes

Alternate Step 3:

```
PROC RANK out=group(keep=grp achieve acedemic) groups=11;
var achieve;
ranks grp;
run;
```

I Doing this in SAS

Step 4: sort the data:

```
PROC SORT data=group1;
  by grp;
```

Step 5: Find sums and means needed:

```
PROC MEANS data=group1 sum mean noprint;
  by grp;
  var academic fitted achieve ;
  output out=grpfit sum=num_aca fit2
    N=num_cases
    mean=dum1 dum2 achbar;
```

I Doing this in SAS

Step 6: Compute various quantities needed for plotting and computing global fit statistics:

```
DATA done;
set grpfit;
p=num_aca/num_cases;
pp = fit2/num_cases;
non_aca = num_cases-num_aca;
fit_non = num_cases-fit2;
Xsq1=((num_aca-fit2)**2)/fit2;
Xsq2=((non_aca-fit_non)**2)/fit_non;
Gsq1= 2* num_aca*log(num_aca/fit2);
Gsq2 = 2*non_aca*log(non_aca/fit_non);
```

Step 7: Sum up the quantities need for the global fit statistics:

```
PROC MEANS data=done sum n ;
var Xsq1 Xsq2 Gsq1 Gsq2;
run;
```

I Results of Last Step...

The MEANS Procedure

Variable	Sum	N
Xsq1	4.3269856	11
Xsq2	5.0778190	11
Gsq1	4.1338700	11
Gsq2	5.5853912	11

$$X^2 = 4.3269856 + 5.0778190 = 9.40$$

$$G^2 = 4.1338700 + 5.5853912 = 9.72$$

$$\begin{aligned} df &= \text{number of logits} - \text{number of parameters} \\ &= 11 - 2 = 9 \end{aligned}$$

I Group Data then Fit Model

Much easier but cruder.

Using the same groups as before, the counts are summed and the model re-fit to the collapsed data. The mean achievement score was used for the numerical value of the explanatory variable.

i.e.,

```
PROC GENMOD data=grpfit;
```

```
    model num_aca/num_cases = achbar / dist=bin link=logit;
```

This yields

$$\text{logit}(\hat{\pi}(x_i)) = -6.9232 + 0.1344x_i$$

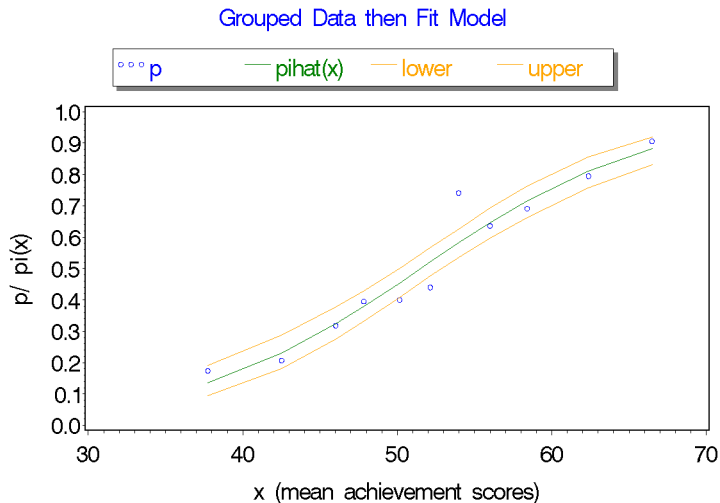
and

Statistic		<i>df</i>	Value	<i>p</i> -value
Deviance	G^2	9	9.7471	.37
Pearson Chi-Square	X^2	9	9.4136	.40

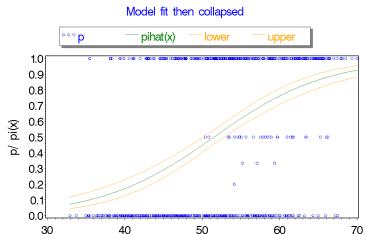
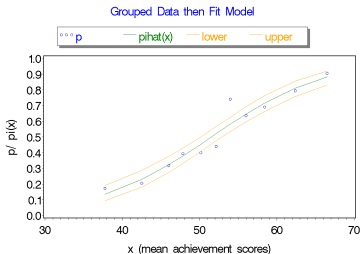
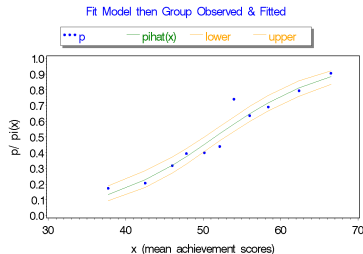
I From Grouping and then Fitting Model

Achievement		Observed Data			
Group	Mean score	# attend academic	# cases in group	Observed proportion	Fitted probability
$x < 40$	37.77	8	46	0.17	.13
$40 \leq x < 45$	42.60	18	87	0.20	.23
$45 \leq x < 47$	46.03	14	44	0.31	.32
$47 \leq x < 49$	47.85	17	43	0.39	.38
$49 \leq x < 51$	50.12	18	50	0.36	.45
$51 \leq x < 53$	52.12	22	50	0.44	.52
$53 \leq x < 55$	53.95	34	58	0.58	.58
$55 \leq x < 57$	56.05	23	44	0.52	.65
$57 \leq x < 60$	58.39	35	68	0.51	.72
$60 \leq x < 65$	62.41	56	78	0.71	.81
$65 \leq x$	66.56	26	32	0.81	.88

I Figure of this



I Comparison



I Alternative Method of Partitioning/Grouping

- **Problem:** When there is more than one explanatory variable, grouping observations becomes more difficult.
- **Solution:** Group the values according to the predicted probabilities such that they have about the same number of observations in each of them.

How:

- 1 Order the observations according to $\hat{\pi}(x)$ from smallest to largest. In the HSB example, there is only 1 explanatory variable and probabilities increase as achievement scores go up (i.e., $\hat{\beta} > 1$) so we can just order the math scores. When there is more than 1 explanatory variable, the ordering must be done using the $\hat{\pi}(x)$'s.
- 2 Depending on the number of groups desired, it is common to partition observations such that they have the same number observations per group.

I Grouping Data by $\hat{\pi}(x_i)$

- (continued) For example, if we wanted 10 groups in the HSB example, then we would try to put $n/10 = 600/10 = 60$ students per group.
The first 60 observations $\rightarrow$ group 1
The next 60 observations $\rightarrow$ group 2
etc.
It's not always possible to have exactly equal numbers in the groups.
- A Pearson-like X^2 computed on the data grouped this way is known as the “Hosmer & Lemeshow” statistic.
It doesn't have a chi-squared distribution, but simulation studies have shown that the distribution of the Hosmer-Lemeshow statistic is approximately chi-squared with $df = g - 2$ (where g = the number of groups).

HSB: Hosmer-Lemeshow statistic = 4.7476, $df = 8$, $p = .7842$.

Conclusion?

I Hosmer-Lemeshow Statistic

PROC/LOGISTIC will compute the Hosmer-Lemeshow statistic, as well as print out the partitioning.

Example using PROC LOGISTIC;

- If data is in individual level format (one line per case):

```
PROC LOGISTIC data=acd descending;
```

```
model academic = achieve / lackfit;
```

```
run;
```

- If data is in tabular form (contingency table):

```
PROC LOGISTIC data=hsbtabs descending;
```

```
model academic/count = achieve / lackfit; run;
```

I Edited Output from PROC LOGISTIC:

The LOGISTIC Procedure Model Information

Data Set	WORK.ACD
Response Variable	academic
Number of Response Levels	2
Number of Observations	600
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered		Total
Value	academic	Frequency
1	1	308
2	0	292

Model Convergence Status

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

I Edited Output from PROC LOGISTIC:

Model Fit Statistics

Criterion	Intercept	Intercept
	Only	and Covariates
AIC	833.350	696.268
SC	837.747	705.062
-2 Log L	831.350	692.268

Testing Global Null Hypothesis: BETA= 0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	139.0819	1	< .0001
Score	127.4700	1	< .0001
Wald	106.4038	1	< .0001

I Edited Output from PROC LOGISTIC:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr >ChiSq
Intercept	1	-7.0543	0.6948	103.0970	< .0001
achieve	1	0.1369	0.0133	106.4038	< .0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
achieve	1.147	1.117 1.177

I Edited Output from PROC LOGISTIC:

Partition for the Hosmer and Lemeshow Test

academic = 1			academic = 0		
Group	Total	Observed	Expected	Observed	Expected
1	60	9	8.67	51	51.33
2	60	13	13.66	47	46.34
3	60	19	18.80	41	41.20
4	60	24	23.74	36	36.26
5	60	26	29.29	34	30.71
6	60	36	33.75	24	26.25
7	60	42	38.06	18	21.94
8	62	41	44.23	21	17.77
9	60	45	47.47	15	12.53
10	58	53	50.34	5	7.66

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
4.7476	8	0.7842

R for Hosmer-Lemeshow:

```
library(ResourceSelection)
```

```
(hs ← hoslem.test(logit.mod$y, fitted(logit.mod), g=10))
```

Hosmer and Lemeshow goodness of fit (GOF) test

```
data: logit.mod$y, fitted(logit.mod)
```

```
X-squared = 3.9814, df = 8, p-value = 0.8588
```

```
names(hs)
```

```
"statistic" "parameter" "p.value" "method" "data.name"
```

```
"observed" "expected"
```

To get information on page ~ 84 of the notes, use `hs$observed` and `hsb$expected`.

I Comparison Tests as Goodness of fit tests

with continuous predictors

More complex models can be fit, such as:

- additional explanatory variables.
- non-linear terms (e.g., x^2).
- interactions
- etc.

and a likelihood ratio test used to compare the model with respect to the more complex models.

If the more complex models do not fit significantly better than the model's fit, then this indicates that the fitted model is reasonable.

Global goodness of fit statistics only indicate that the model does not fit perfectly (i.e., there is some lack of fit). By comparing a model's fit with more complex models provides test for particular types of lack of fit.

I Likelihood-Ratio Model Comparison Tests

The likelihood-ratio statistic equals

$$\text{Likelihood-ratio statistic} = -2(L_0 - L_1)$$

where

- L_1 = the maximized log of the likelihood function from a complex model, say M_1 .
- L_0 = the maximized log of the likelihood function from a simpler (nested) model, say M_0 .

The goodness of model fit statistic G^2 is a special case of the likelihood ratio test statistic where

- $M_O = M$, the model we're testing.
- $M_1 = M_S$, the most complex model possible or the “saturated” model.

For Poisson and logistic regression, G^2 is equal to “deviance” of the model.

I Likelihood-Ratio Model Comparison Tests

- L_S = maximized log of the likelihood function for the saturated model.
- L_O = maximized log of the likelihood function for the simpler model M_0 .
- L_1 = maximized log of the likelihood function for the complex model M_1 .

where we want to compare the fit of the model M_0 and M_1 .

$$\text{deviance for } M_0 = G^2(M_0) = -2(L_0 - L_S)$$

and

$$\text{deviance for } M_1 = G^2(M_1) = -2(L_1 - L_S)$$

The likelihood ratio statistic

$$\begin{aligned} G^2(M_0|M_1) &= -2(L_0 - L_1) \\ &= -2[(L_0 - L_S) - (L_1 - L_S)] \\ &= G^2(M_0) - G^2(M_1) \end{aligned}$$

I and ...

and $df = df_0 - df_1$.

Assuming that M_1 holds, this statistic tests

- Whether the lack of fit of M_0 is significantly larger than that of M_1 .
- Whether the parameters in M_1 that are not in M_0 equal zero.

HSB example using the grouped data:

- M_0 = Model with only an intercept

$$\text{logit}(x_i) = \alpha$$

$$G^2(M_0) = 144.3546 \text{ with } df_0 = (11 - 1) = 10$$

- M_1 = Model with an intercept and math scores

$$\text{logit}(x_i) = \alpha + \beta x_i$$

$$G^2(M_1) = 12.76 \text{ with } df_1 = (11 - 2) = 9.7471.$$

- $G^2(M_0|M_1) = 144.3546 - 9.7471 = 134.61$ with $df = 10 - 9 = 1$ and $p\text{-value} < .0001$.

I The easier/quicker way

Or using the type3 option in the GENMOD MODEL statement:

LR Statistics For Type 3 Analysis

		Chi-	
Source	DF	Square	Pr > ChiSq
ach_bar	1	134.57	<.0001

In R: `anova(logit.mod)` and then use `1-pchisq(...value of statistic...,1)`.

I Summary Measures of Predictive Power

Receiver Operating Characteristic (ROC), Classification tables, and the Concordance index.

Suppose we have a simple model

$$\text{logit}(\hat{\pi}_i) = \hat{\alpha} + \hat{\beta}x_i$$

Let π_o be a cut-point or cut-score and $\hat{\pi}_i$ be a predicted probability of the model. The predicted response is

$$\hat{y}_i = \begin{cases} 1 & \text{if } \hat{\pi}_i > \pi_o \\ 0 & \text{if } \hat{\pi}_i \leq \pi_o \end{cases}$$

Classification Table:

		Predicted	
		$\hat{y}_i = 1$	$\hat{y}_i = 0$
Actual	$y = 1$	correct	incorrect (false negative)
	$y = 0$	incorrect (false positive)	correct

I Classification Table

We're more interested in conditional proportions and probabilities:

		Predicted		
		$\hat{y}_i = 1$	$\hat{y}_i = 0$	
Actual	$y = 1$	n_{11}/n_{1+}	n_{12}/n_{1+}	n_{1+}
	$y = 0$	n_{21}/n_{2+}	n_{22}/n_{2+}	n_{2+}

$$n_{11}/n_{1+} = \text{proportion } (\hat{y} = 1|y = 1) = \text{"sensitivity"}$$

$$n_{22}/n_{2+} = \text{proportion } (\hat{y} = 0|y = 0) = \text{"specificity"}$$

$$\begin{aligned}
 p(\text{correct}) &= p(\hat{y} = 1 \ \& \ y = 1) + p(\hat{y} = 0 \ \& \ y = 0) \\
 &= p(\hat{y} = 1|y = 1)p(y = 1) + (\hat{y} = 0|y = 0)p(y = 0) \\
 &= (\text{sensitivity})p(y = 1) + (\text{specificity})p(y = 0)
 \end{aligned}$$

I HSB Example

- Let the cut-score equal $\pi_o = .50$.
- Compare $\hat{\pi}_i$ and classify i as $Y = 1$ if $\hat{\pi}_i > \hat{\pi}_o$, otherwise classify i as $Y = 0$.
- Tabulate the results

		Predicted		
		$\hat{y} = 1$	$\hat{y} = 0$	
Actual	$y = 1$	227	81	308
	$y = 0$	94	198	292

- The Conditional proportions:

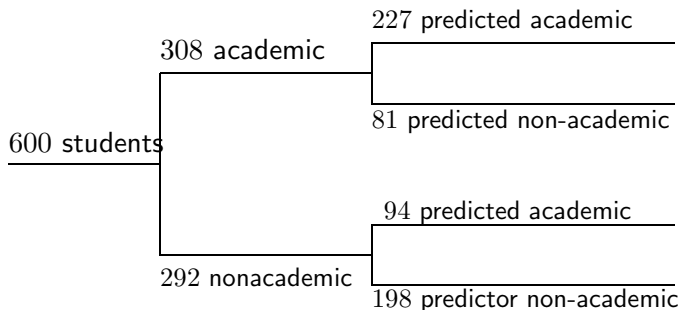
$$\text{Sensitivity} = 227/308 = .737$$

$$\text{Specificity} = 198/292 = .678$$

- The proportion correct
 $= .708(308/600) + .678(292/600) = .5387 + .3481 = .71$

I HSB Example with $\pi_o = .50$

$$\text{Sensitivity} = 227/308 \times 100\% = 73.7\%$$



$$\text{Specificity} = 198/292 \times 100\% = 67.8\%$$

$$\text{Percent correct} = (227 + 198)/600\% = 425/600 \times 100\% = 70.8\%$$

I Sensitivity, Specificity & p(Correct)

- For every cut-score you will get a different result.
- For HSB, using a cut-score of $\pi_o = .70$ yields

		Predicted		
		$\hat{y} = 1$	$\hat{y} = 0$	
Actual	$y = 1$	112	180	308
	$y = 0$	30	278	292

- $$\text{Sensitivity} = 112/308 = .384$$

$$\text{Specificity} = 278/292 = .903$$

$$\text{Correct} = (112 + 278)/600 = .650$$

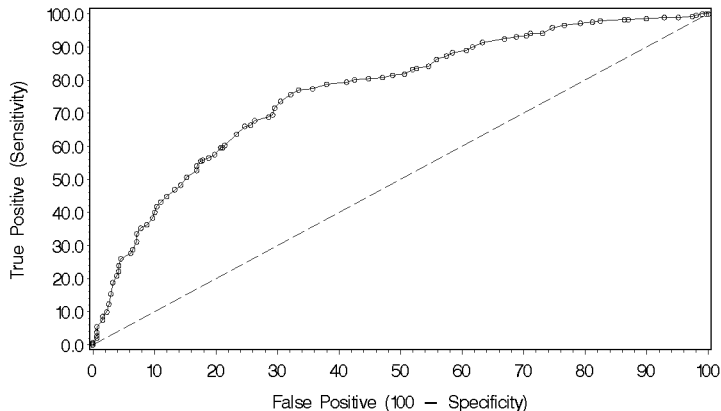
- Do this for lots of possible cut-scores and plot the results $\rightarrow$ ROC curve.

I ROC Curve for HSB: $c = .769$

Cutpoint .0 to 1.0 by .01

Sensitivity = Percent Correctly Classified as Academic

100 - Specificity = Percent Incorrectly Classified as Academic



I Area Under ROC Curve

Concordance: Take two cases i and j where $y_i = 1$ and $y_j = 0$ ($i \neq j$),

If $\hat{\pi}_i > \hat{\pi}_j$, then the pair is concordant

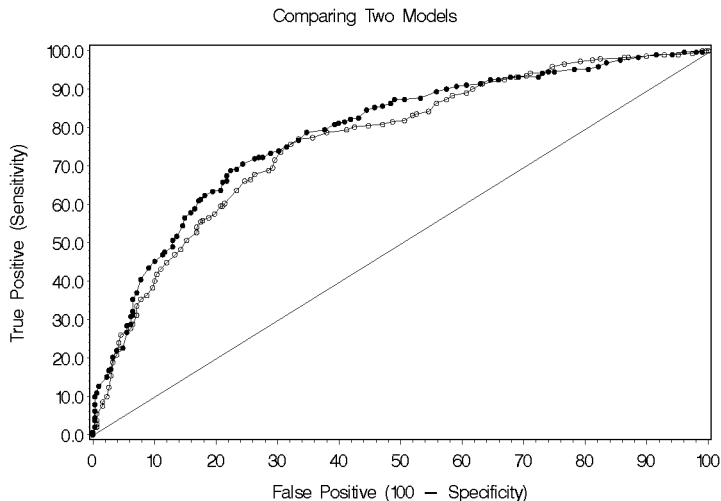
If $\hat{\pi}_i < \hat{\pi}_j$, then the pair is discordant

If $\hat{\pi}_i = \hat{\pi}_j$, then the pair is tie

The area under the ROC curve equals the concordance index. The concordance index is an estimate of the probability that predictions and outcomes are concordant. In PROC LOGISTIC, this index is c in the table of “Association of Predicted Probabilities and Observed Responses”.

This also provides a way to compare models, the solid dots in the next figure are from a model with more predictors. For example,...

I ROC Curve for HSB: $c = .769$ and $.809$



I SAS: Concordance and ROC Curve

Below is a “bare-bones” code to obtain concordance index and plot of ROC Curve. For more on plotting (eg., 2 curves), see SAS that goes with lecture notes.

```
/* Simple ROC plot */
proc logistic data = hsb;
model academic= achieve / outroc =ROCDData;
run;

symbol1 v=dot i=join;
proc gplot data=ROCDData;
plot _sensit_ * _1mspec_;
run;
```

I R: Concordance and ROC Curve

```
library(DescTools)
# Cocordance index
Cstat(logit.mod,resp=hsb$program)

# Plot of ROC curve
# Uses package pROC
roc1 ← roc(hsb$program ~ hsb$sachieve, auc=TRUE, ci=TRUE,
plot=TRUE)
names(roc1)

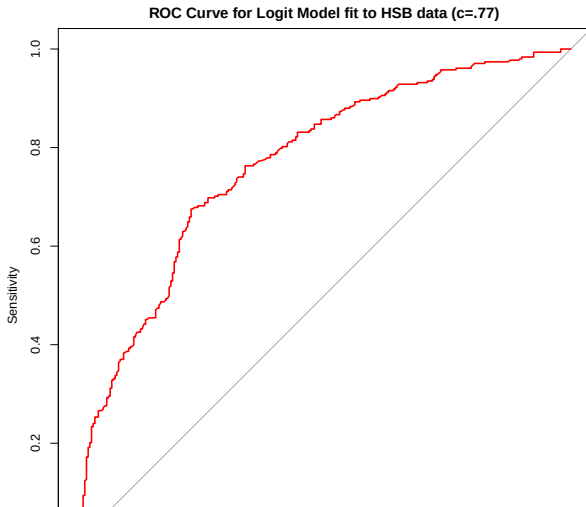
# -- area under curve (i.e. concordance)
roc1$auc

# -- concordance index (area under curve) and confidence
interval for it.
roc1$ci
```

I R: Concordance and ROC Curve

```
# -- a 2nd ROC plot -- touched up a bit
plot(roc1,
main="ROC Curve for Logit Model fit to HSB data (c=.77)",
r col="red")
```

I R: Concordance and ROC Curve



I Residuals

Goodness-of fit-statistics are global, summary measures of lack of fit.
We should also

- Describe the lack of fit
- Look for patterns in lack of fit.

Let

- y_i = observed number of events/successes.
- n_i = number of observations with explanatory variable equal to x_i .
- $\hat{\pi}_i$ = the estimated (fitted) probability at x_i .
- $n_i\hat{\pi}_i$ = estimated (fitted) number of events.

Pearson residuals are

$$e_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1 - \hat{\pi}_i)}}$$

I Pearson Residuals

Pearson residuals are

$$e_i = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}}$$

- $X^2 = \sum_i e_i^2$; that is, the e_i 's are components of X^2 .
- When the model holds, the Pearson residuals are approximately normal with mean 0 and variance slightly less than 1. Values larger than 2 are “large”.
- Just as X^2 and G^2 are not valid when fitted values are small, Pearson residuals aren't that useful (i.e., they have limited meaning).
If $n_i = 1$ at many values, then the possible values for $y_i = 1$ and 0, so e_i can assume only two values.

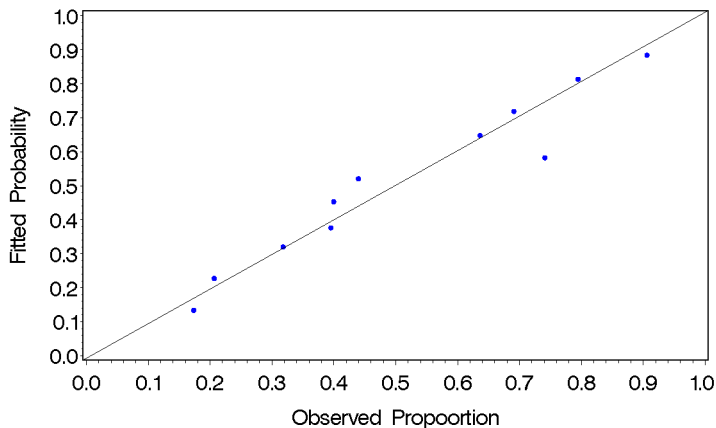
I HSB Using Grouped Data

(note: Computing residuals on un-grouped data not useful due to small n per “cell”).

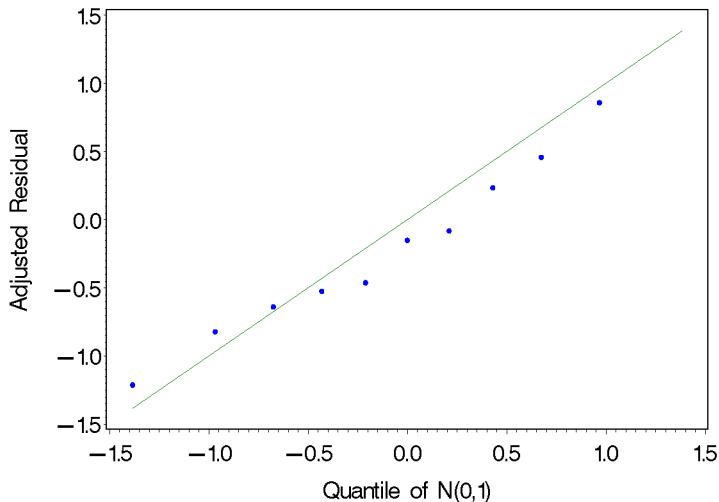
Group	mean achieve	# attend acad	Number of cases	Observed prop	Predicted prob	Pearson residuals
1	37.67	8	46	.17	.13	.78
2	42.60	18	87	.21	.23	-.55
3	46.01	14	44	.32	.32	.07
4	47.87	17	43	.40	.38	.21
5	50.11	20	50	.40	.45	-.75
6	52.17	22	50	.44	.52	-1.15
7	53.98	43	58	.74	.58	2.47
8	56.98	28	44	.63	.65	-.15
9	58.42	47	68	.69	.72	-.45
10	62.43	62	78	.79	.81	.38
11	66.56	29	32	.90	.88	.42

I Observed versus Fitted (Grouped Data)

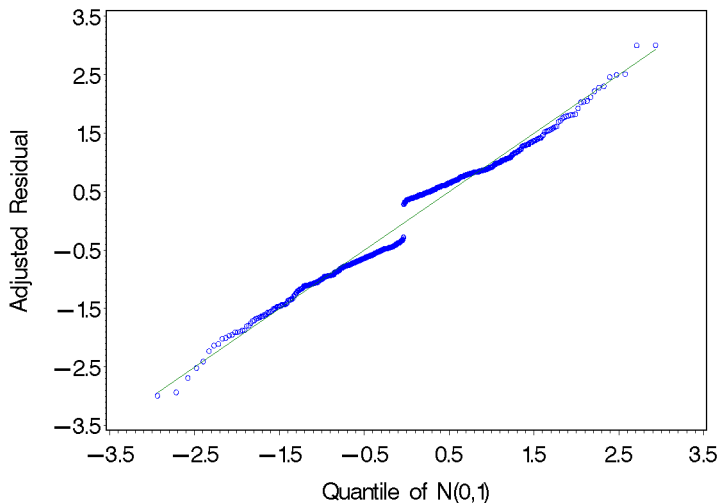
Fit Model then Group Observed & Fitted



I Q-Q Plot of Pearson Residuals (Grouped)



I Q-Q Plot of Pearson Residuals (not grouped)



I Getting Data for Q-Q Plot

- **Step1:** Fit model and save residuals to SAS file:

```
PROC GENMOD data=hsb;
    model academic/ncases = achieve / link=logit
        dist=binomial obstats type3 covb;
    output out=preds pred =fitted stdreschi=adjusted;
```
- **Step 2:** Create a SAS data file with quantiles from normal distribution:

```
DATA QQprep;
    do p=1 to 600;
        prop=p/601;
        z = quantile('normal',prop);
        output;
    end;
```

I Hard way to Make Q-Q Plot

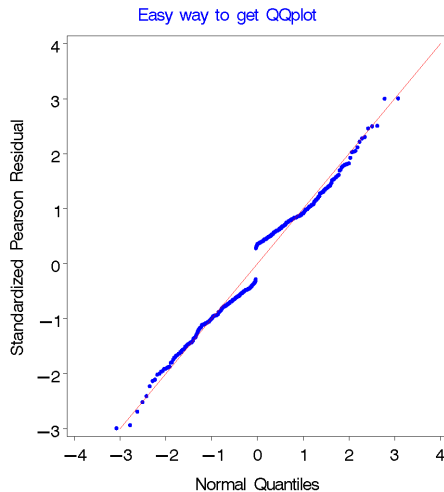
- **Step 3:** Sort data file with residuals by the values of the residuals;
PROC SORT data=preds;
by adjusted;
- **Step 4:** Merge the two files:
DATA QQplot;
merge preds QQprep;

I Easy way to Make Q-Q Plot

```
proc genmod data=hsb;
  model academic/ncases = achieve /link=logit dist=bin;
  output out=outnew pred=grpfit lower=lo
    upper=hi stdreschi=adjusted;
  title 'Easy way to get QQplot';
run;

proc univariate data=outnew;
  var adjusted;
  qqplot / normal(mu=0 sigma=1) square ctext=black ;
run;
```

I Result of Easy Way



I R to Make Q-Q Plot

```
# Pearson (ajdusted) standardized residuals
resid ← rstandard(logit.mod, pearson=TRUE)
```

```
qqnorm(resid,
        main="QQ plot of Pearson Standardized Residuals")
```

I Influence

Some observations may have too much “**influence**” on

- ① Their effect on **parameter estimates** — If the observation is deleted, the values of parameter estimates are considerably different.
- ② Their effect on the **goodness-of-fit** of the model to data — If the observation is deleted, the change in how well the model fits the data is large.
- ③ The effect of coding or **misclassification error** of the binary response variable on statistic(s) of interest. Statistics of interest include fit statistics and/or model parameter estimates.

I Measures for Detecting Influence

They are primarily designed to detect one or the other of these two aspects.

The first three types:

- Often influential observations are those that are extreme in terms of their value(s) on the explanatory variable(s).
- Are pretty much generalizations of regression diagnostics for normal linear regression.
- “range of influence” are designed specifically for logistic regression for binary responses.

Additional references:

- Collett, D.R. (1991). *Modelling Binary Data*. London: Chapman & Hall.
- Pregibon, D (1981). Logistic regression diagnostics. *Annals of Statistics*, 9, 705–724.
- Hosmer, D.W. & Lemeshow, S. (1989). *Applied Logistic Regression*. New York: Wiley.

I The Measures for Detecting Influence

- “Leverage” (diagonal elements of the *hat* matrix).
- Pearson, deviance, and adjusted residuals.
- *Dfbeta*.
- *c* and $\bar{c}$.
- Change in X^2 or G^2 goodness-of-fit statistics (i.e., *DIFCHISQ* and *DIFDEV*, respectively).
- Range of Influence (ROI) statistics.

Each of these measures

- Are computed for each observation.
- The larger the value, the greater the observation's influence.
- All are computed by PROC LOGISTIC, except the adjusted residuals (need to use PROC GENMOD) and range frequency statistics (I wrote a set of SAS MACROs for this).
- I wrote an R function called “roi”.

I Example Data: ESR

Number of cases (people) = 32

Response variable is whether a person is healthy or not (based on ESR).

Model probability that a person is healthy as a function of

- FIBRIN: level of plasma fibrinogen.
- GLOBULIN: level of gamma-globulin.

The model with both explanatory variables:

- Test statistics for GLOBULIN are not significant ($df = 1$).
Wald = 1.698 (from model with 2 predictors)
Likelihood ratio = $24.840 - 22.971 = 1.87$
- Test statistics for FIBRIN are significant ($df = 1$).
Wald = 3.87 ($p = .05$) (from model with 2 predictors)
Likelihood ratio = $28.945 - 22.971 = 5.974$ ($p = .01$)

I Simpler Model for ESR

The model with just FIBRIN:

Test statistics for FIBRIN are significant ($df = 1$):

Wald = 4.11 ($p = .04$)

Likelihood ratio = 5.974 ($p = .01$)

Hosmer & Lemeshow statistic = 10.832 with $df = 8$ & $p = .21$.

Concordance index = .71.

I ESR Data & SAS

* fibrin = level of plasma fibrinogen (gm/liter)
 globulin = level of gamma-globulin (gm/liter)
 response = (0 esr<20 or unhealthy, 1 esr>=20 for healthy)
 where esr=erythrocyte sedimentation rate;

data esr;

title'ESR Data';

input id fibrin globulin response @@;

n=1;

datalines;

1	2.52	38	0	2	2.56	31	0	3	2.19	33	0	4	2.18	31	0
5	3.41	37	0	6	2.46	36	0	7	3.22	38	0	8	2.21	37	0
9	3.15	39	0	10	2.60	41	0	11	2.29	36	0	12	2.35	29	0
13	5.06	37	1	14	3.34	32	1	15	2.38	37	1	16	3.15	36	0
17	3.53	46	1	18	2.68	34	0	19	2.60	38	0	20	2.23	37	0
21	2.88	30	0	22	2.65	46	0	23	2.09	44	1	24	2.28	36	0
25	2.67	39	0	26	2.29	31	0	27	2.15	31	0	28	2.54	28	0
29	3.93	32	1	30	3.34	30	0	31	2.99	36	0	32	3.32	35	0

I ESR Data & SAS

```

/* Example 1 */ ;
proc logistic data=esr descending;
model response=globulin;
proc logistic data=esr descending;
model response=fibrin globulin;
title 'ESR Data';
run;

/* Example 2 */
ods html;
ods graphics on;
proc logistic data=esr descending;
model response=fibrin / lackfit influence iplots;
title 'ERS Data';
run;
ods graphics off;
ods html close;
run;

```

I ESR Data & SAS

```

data esr2;
input fibrin response n @@;
datalines;
2.09 1 1 2.15 0 1 2.18 0 1 2.19 0 1
2.21 0 1 2.23 0 1 2.28 0 1 2.29 0 2
2.35 0 1 2.38 1 1 2.46 0 1 2.52 0 1
2.54 0 1 2.56 0 1 2.60 0 2 2.65 0 1
2.67 0 1 2.68 0 1 2.88 0 1 2.99 0 1
3.15 0 2 3.22 0 1 3.32 0 1 3.34 1 2
3.41 0 1 3.53 1 1 3.93 1 1 5.06 1 1
proc genmod order=data;
  model response/n = fibrin /link=logit dist=binomial obstats residuals
  type3;
  title 'GENMOD: Logistic regression with fibrin';

```

I ESR Data & R

```
esr <- read.table(file="esr_data.txt", header=TRUE)
```

```
head(esr)
```

	id	fibrin	globulin	response
1	1	2.52	38	0
2	2	2.56	31	0
3	3	2.19	33	0
4	4	2.18	31	0
5	5	3.41	37	0
	:			

```
model.f <- glm(response ~ fibrin , data=esr,
family=binomial)
```

I ESR Data & R: Influence measures

Influence measures of

```
glm(formula = response ~ fibrin, family = binomial, data = esr) :
(infl.f ← influence.measures(model.f))
```

	dfb.1	dfb.fbrn	dffit	cov.r	cook.d	hat inf
1	-0.08085	0.06781	-0.1025	1.097	0.00235	0.0406
2	-0.08052	0.06652	-0.1055	1.095	0.00249	0.0402
3	-0.07215	0.06505	-0.0788	1.107	0.00136	0.0429
⋮						

I 1. Leverage or h_i

These equal the diagonal elements of the “hat” matrix, which has a row and column corresponding to each observation.

- The hat matrix is applied to sample logits yields the predicted logits for the model.
- h_i is good for detecting extreme points in the design space.
- Qualification:
 - The more extreme the estimated probability (i.e., $\hat{\pi}(x) < .1$ or $\hat{\pi}(x) > .9$), the smaller the h_i .
 - Therefore, when an observation has a very small or very large estimated probability, h_i is not a good detector of the observation's distance from the design space.

I 2. Pearson, Deviance, and ...

Used to identify observations that are not explained very well by model.

- Pearson residuals:

$$e_i = \frac{y_i - n_i \hat{\pi}(x_i)}{\sqrt{n_i \hat{\pi}(x_i)(1 - \hat{\pi}(x_i))}}$$

- Deviance residuals (where $\hat{\pi}(x_i) = \hat{\pi}_i$)

$$\begin{aligned} d_i &= -\sqrt{-2n_i \log(1 - \hat{\pi}_i)} && \text{if } y_i = 0 \\ &= -\sqrt{2\{y_i \log(y_i/(n_i \hat{\pi}_i)) + (n_i - y_i) \log((n_i - y_i)/(n_i(1 - \hat{\pi}_i)))\}} && \text{if } y_i/n_i < \hat{\pi}_i \\ &= +\sqrt{2\{y_i \log(y_i/(n_i \hat{\pi}_i)) + (n_i - y_i) \log((n_i - y_i)/(n_i(1 - \hat{\pi}_i)))\}} && \text{if } y_i/n_i > \hat{\pi}_i \\ &= \sqrt{-2n_i \log(\hat{\pi}_i)} && \text{if } y_i = n_i \end{aligned}$$

I ...and adjusted Residuals

Adjusted Residuals are Pearson residuals divided by $(1 - h_i)^{1/2}$:

$$\text{Adjusted residual} = \frac{e_i}{\sqrt{1 - h_i}} = \frac{y_i - n_i \hat{\pi}(x_i)}{\sqrt{n_i \hat{\pi}(x_i)(1 - \hat{\pi}(x_i))(1 - h_i)}}$$

I 3. *Dfbeta*

This assesses the effect that an individual observation has on the parameter estimates.

$$Dfbeta = \frac{\text{change in parameter estimate}}{\text{standard error of change}}$$

- The larger the value of *Dfbeta*, the larger the change in the estimated parameter when the observation is removed.
- Large value indicates that certain observations are leading to instability in the parameter estimates.
- *PROC LOGISTIC* uses a 1 step method to approximate *Dfbeta*.

I 4. c and $\bar{c}$

These measure the change in the joint confidence interval for the parameters produced by deleting an observation.

- These use the same idea as “Cook distances” in ordinary linear regression.
- *PROC LOGISTIC* uses a 1 step method to approximate them

$$c_i = \frac{e_i^2 h_i}{(1 - h_i)^2}$$

and

$$\bar{c}_i = \frac{e_i h_i}{(1 - h_i)}$$

- In using these statistics, it is useful to plot them versus some index (e.g., observation number).

I 5. *DIFCHISQ* and *DIFDEV*

Equal the change in the X^2 and G^2 goodness of fit statistics for the model when the observation is deleted.

They are diagnostics for detecting observations that contribute heavily to the lack of fit of the model to the data.

With a large number of observations, the time that it would take to actually delete each observation and fit the model to obtain the actual change in X^2 and G^2 could be prohibitive, so SAS/LOGISTIC uses a 1 step method to estimate the change.

- *PROC LOGISTIC* uses a 1 step method to estimate the change in X^2 (i.e., *DIFCHISQ*):

$$DIFCHISQ = \frac{\bar{c}_i}{h_i}$$

- *PROC LOGISTIC* uses a 1 step method to estimate the change in G^2 (i.e., *DIFDEV*):

$$DIFDEV = d_i^2 + \bar{c}_i$$

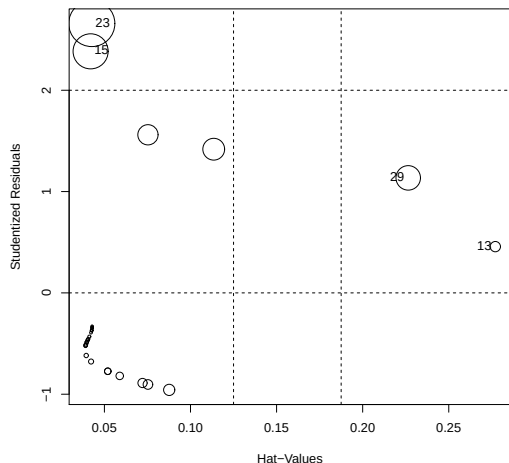
I Example: ESR data

For code for SAS PROC LOGISTIC and R, see course web-site examples.
Following is some output from R.

In console after getting graph your all gets list of potentially influential observations

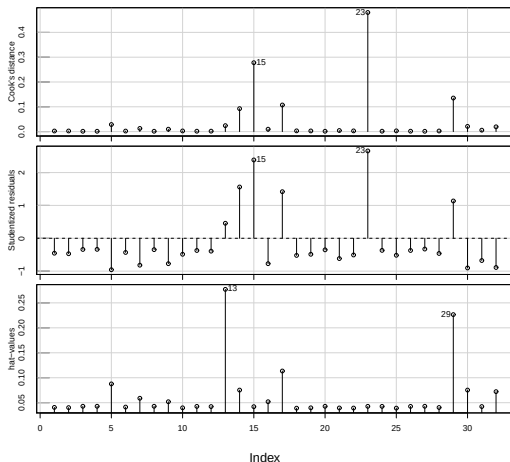
	StudRes	Hat	CookD
13	0.4565581	0.27704251	0.02404385
15	2.3840402	0.04198596	0.27771524
23	2.6584101	0.04268015	0.48024585
29	1.1350206	0.22645872	0.13531633

I Studentized Residuals by Hat-Values

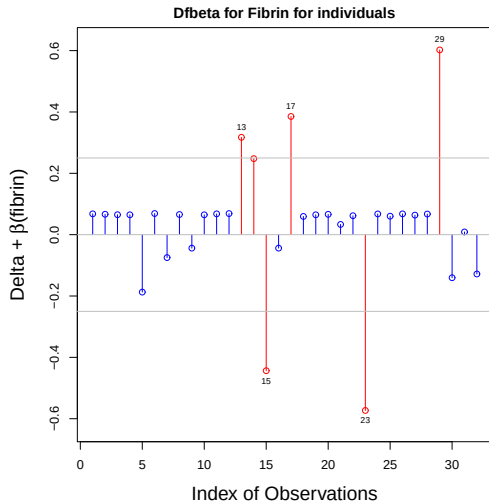


I Cook, Studentized Residuals, Hat-Values

Diagnostic Plots



I Dfbeta: plasma fibrinogen



I Example: ESR data — summary (so far)

X = an extreme value and **x** = a noticeably different value

		Case Number					
Diagnostic measure		13	14	15	17	23	29
Leverage	h_i	X			X		X
Pearson residual	e_i	x	x	X	x	X	x
Deviance residual	d_i	x	X	X	X	X	X
Change in parameter estimate	α (intercept)			X		X	x
	β (fibrin)	x	x	X	x	X	x
Change in CI	c_i		x	X	x	X	x
	$\bar{c}_i$		x	X	x	X	x
Change in X^2			x	X	x	X	x
Change in G^2			x	X	x	X	x

- Cases 15 & 23 appear to be influential.
- There were 6 cases that were classified as unhealthy: 13, 14, 15, 17, 23, 29

I Range of Influence (ROI) Statistic

Reference: Fay, M.P. (2002). Measuring a binary response's range of influence in logistic regression. *American Statistician*, 56, 5–9.

- Purpose/Problem:

- There is always the possibility that there was a misclassification on the response variable or a data entry error in the response variable.
- If it is difficult to check the classification (time consuming and/or expensive), you would like to identify a sub-set of “questionable” cases.

- Solution: Range of Influence Statistic.

I Computing ROI Statistic

- ① Fit the logistic regression model using data (as is).
- ② For each case, change the value of the (binary) response variable,

$$y_i^* = 1 - y_i,$$

and re-fit the logistic regression model.

- ③ Compute the difference between statistics using the changed data and the un-altered data, which is called the “Range of Influence Statistic.”
- ④ Look for cases that have extreme values.
- ⑤ I wrote SAS MACRO to compute ROI statistics and wrote a function for R.

I ROI Statistic: ESR data

Model that we settled on was

$$\begin{aligned}\text{logit}(\hat{\pi}_i) &= \hat{\alpha} + \hat{\beta}(\text{fibrin})_i \\ &= -6.8451 + 1.8271(\text{fibrin})_i\end{aligned}$$

Which has $\ln(\text{likelihood}) = -12.4202$.

We'll look at ROI statistics for

- Intercept:

$$\Delta(\alpha)_i = \hat{\alpha}_i^* - \hat{\alpha} = \hat{\alpha}_i^* - (-6.8451)$$

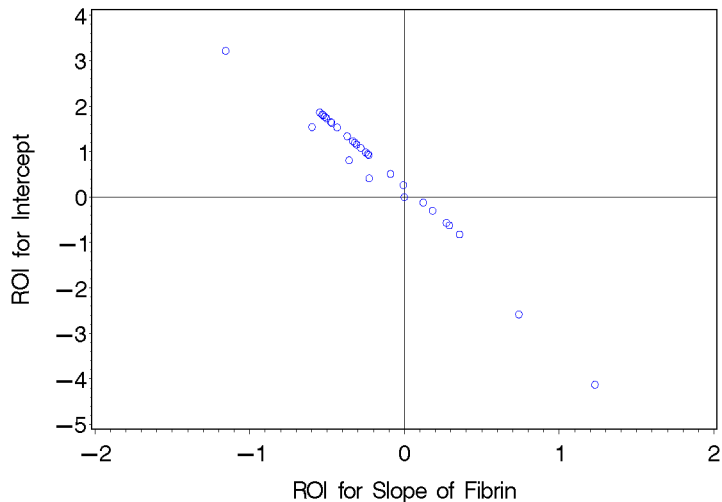
- Slope for fibrin:

$$\Delta(\beta)_i = \hat{\beta}_i^* - \hat{\beta} = \hat{\beta}_i^* - 1.8271$$

- Goodness-of-fit:

$$\Delta(\text{lnlike})_i = \ln(\text{likelihood})_i^* - \ln(\text{likelihood}) = \ln(\text{likelihood})_i^* - 12.4202$$

I Plot of ROI for ESR data



I List of ROI for ESR data

ESR Data

Obs	obs changed	delta a	delta b	delta ll	
1	13	3.22002	-1.15387	-0.98268	**
2	29	1.54184	-0.59611	0.08378	
3	27	1.86539	-0.54622	-2.54509	
4	4	1.81739	-0.52996	-2.50642	
5	3	1.80128	-0.52450	-2.49342	
6	8	1.76887	-0.51353	-2.46726	
7	20	1.73622	-0.50247	-2.44087	
8	24	1.65351	-0.47447	-2.37395	
9	11	1.63678	-0.46881	-2.36039	
10	26	1.63678	-0.46881	-2.36039	
11	12	1.53500	-0.43437	-2.27786	
12	6	1.34190	-0.36908	-2.12110	
13	17	0.81094	-0.35647	0.62283	
14	1	1.23279	-0.33221	-2.03251	
15	28	1.19579	-0.31971	-2.00249	
16	2	1.15846	-0.30710	-1.97222	

I List of ROI for ESR data (continued)

ESR Data					
Obs	obs changed	delta a	delta b	delta ll	
17	10	1.08282	-0.28156	-1.91091	
18	19	1.08282	-0.28156	-1.91091	
19	22	0.98634	-0.24900	-1.83282	
20	25	0.94714	-0.23577	-1.80112	
21	18	0.92740	-0.22911	-1.78518	
22	14	0.41425	-0.22689	0.91436	
23	21	0.51242	-0.08926	-1.45180	
24	31	0.26625	-0.00644	-1.25611	
25	0	0.00003	0.00000	0.00002	
26	16	-0.11873	0.12280	-0.95460	
27	9	-0.11873	0.12280	-0.95460	
28	7	-0.29759	0.18275	-0.81602	
29	32	-0.56549	0.27242	-0.61059	
30	30	-0.62092	0.29095	-0.56842	
31	5	-0.82005	0.35747	-0.41784	
32	15	-2.57889	0.74032	2.85995	***
33	23	-4.12667	1.23224	3.67215	***

I SAS & Computing ROI

SAS/MACRO

```
%RangeOfInfluence(indata=esr,rangedata=range,samplesize=32,
  obs=response,Y=response,linpred=fibrin,discrete=);
```

where

- The response variable should equal 0/1
- In main macro “%rangeinfluence”
 - indata= the data set that is being analyzed
 - rangedata= data set that is output that included parameter estimates and loglike for complete data set (1st line) and changing the response of each of the lines of data (one at a time).
Use this data set to compute desired range of influence statistics.
 - samplesize= number of observations
 - obs= name of the identification variable (like an id). Macro assumes this equals 1 - sample size
 - Y = name of the response variable
 - linpred = list of variables in the linear predictor
 - discrete = list of discrete (nominal) variables in the linear predictor

I Running %RangeOfInfluence

```
title 'Example 1: One numerical explanatory variable';
%RangeOfInfluence(indata=esr, rangedata=range1,
    samplesize=32, obs=id,
    Y=response,
    linpred=fibrin );
run;
```

```
title 'Example 2: Two numerical explanatory variables';
%RangeOfInfluence(indata=esr, rangedata=range2,
    samplesize=32, obs=id,
    Y=response,
    linpred=fibrin globulin );
run;
```

I Running %RangeOfInfluence

```

title 'Example 3: discrete variable  numerical predictors';
%RangeOfInfluence(indata=esr2, rangedata=range3,
    samplesize=32, obs=id,
    Y=response,
    linpred=fibrin egdiscrete,
    discrete=egdiscrete );
run;

```

I After Running %RangeOfInfluence

```

data range;
set range1;
  if obs_changed < 1 then delete;
  delta_a = intercept - ( -6.8451);
  delta_b = fibrin - ( 1.82708 );
  delta_ll= _ LNLIKE _ - (-12.4202);
run;
proc sort data=rangestat;
  by delta_b;
proc print;
  var obs_changed delta_a delta_b delta_ll;
run;

```

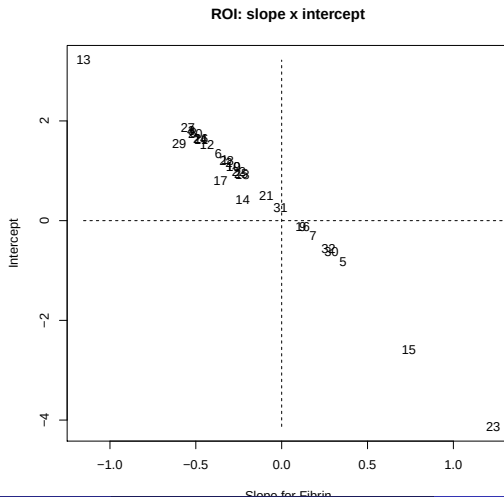
I R function: roi

- Download the function from the course web-site.
- Enter function in the console window (minus the examples at the end)
- You need to change the response variable to “y” and specify a model using the formula command; that is,

```
esr$y ← esr$response
model1 ← formula(y ~ fibrin)
roi.list ← roi(esr,model1)
```

- Take a look at the values in roi.list.

I Plot using roi data



I Summary for ESR Data

The model with just FIBRIN.

- Test statistics for coefficient of FIBRIN is significant ($df = 1$)
Wald statistic = 4.11 (p -value = .04)
Likelihood ratio = 6.04 (p -value = .01)
- Hosmer & Lemeshow statistic = 10.832 with $df = 8$ and p -value = .21.
- From the regression diagnostics
 - Case number 15 and 23 appear to be influential.
 - The 6 who were unhealthy are 13, 14, 15, 17, 23, and 29.

I 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.

I Titanic Data Set

Data can be found on course web-site and online
For more information, google “Titanic data set”

Data Available:

- 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.

...but first we need to discuss qualitative explanatory variables & multiple explanatory variables...