Chapter 5: Logistic Regression-II

Dipankar Bandyopadhyay

Department of Biostatistics,
Virginia Commonwealth University

BIOS 625: Categorical Data & GLM

[Acknowledgements to Tim Hanson and Haitao Chu]

Alcohol consumption and malformation example continued:

- Let's use $X = 1$ as the reference level. Then the model is

  $$\text{logit}\,\pi(X) = \beta_0 + \beta_1 I\{X=2\} + \beta_2 I\{X=3\} + \beta_3 I\{X=4\} + \beta_4 I\{X=5\}.$$

- We may be interested in the how the odds of malformation changes when dropping from 3-4 drinks per week ($X = 4$) to less than one drink per week ($X = 2$), given by $e^{\beta_3 - \beta_1}$.
- A *contrast* is a linear combination $\mathbf{c}'\boldsymbol{\beta} = c_1\beta_1 + c_2\beta_2 + \cdots + c_{p-1}\beta_{p-1}$. We are specifically interested in $H_0 : \beta_3 = \beta_1$, or equivalently, $H_0 : \beta_3 - \beta_1 = 0$, as well as estimating $e^{\beta_3 - \beta_1}$.

```
proc logistic data=mal;
    class cons / param=ref ref=first ;
    model present/total = cons;
    contrast "exp(b3-b1)" cons -1 0 1 0 / estimate=exp;
    contrast "b3-b1" cons -1 0 1 0 / estimate;
run;
```

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi−Square	Pr > ChiSq
Intercept		1	−5.8736	0.1445	1651.3399	<.0001
cons	2	1	−0.0682	0.2174	0.0984	0.7538
cons	3	1	0.8136	0.4713	2.9795	0.0843
cons	4	1	1.0374	1.0143	1.0460	0.3064
cons	5	1	2.2632	1.0235	4.8900	0.0270

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
cons 2 vs 1	0.934	0.610	1.430
cons 3 vs 1	2.256	0.896	5.683
cons 4 vs 1	2.822	0.386	20.602
cons 5 vs 1	9.614	1.293	71.460

- Let θ_{ij} be the odds ratio for malformation when going from level $X = i$ to $X = j$.
- We automatically get $\hat{\theta}_{21} = e^{-0.068} = 0.934$, $\hat{\theta}_{31} = e^{0.814} = 2.26$, etc.
- Because $\theta_{42} = \theta_{41}/\theta_{21}$ we can estimate $\hat{\theta}_{42} = 2.822/0.934 = 3.02$, or else directly from the dummy variable coefficients, $e^{1.037-(-0.068)} = 3.02$.
- The CONTRAST command allows us to further test $H_0 : \beta_3 = \beta_1$ and to get a 95% CI for the odds ratio $\theta_{42} = e^{\beta_3 - \beta_1}$.

Contrast Test Results

Contrast	DF	Wald Chi−Square	Pr > ChiSq
exp(b3−b1)	1	1.1817	0.2770
b3−b1	1	1.1817	0.2770

Contrast Rows Estimation and Testing Results

Contrast	Type	Row	Estimate	Standard Error	Alpha	Confidence Limits		Wald Chi−Square	Pr > ChiSq
exp(b3−b1)	EXP	1	3.0209	3.0723	0.05	0.4116	22.1728	1.1817	0.277
b3−b1	PARM	1	1.1056	1.0170	0.05	−0.8878	3.0989	1.1817	0.277

We are allowed linear contrasts or the exponential of linear contrasts. To get, for example, the *relative risk* of malformation,

$$h(\beta) = \frac{P(Y=1|X=4)}{P(Y=1|X=2)} = \frac{e^{\beta_0+\beta_3}/[1+e^{\beta_0+\beta_3}]}{e^{\beta_0+\beta_1}/[1+e^{\beta_0+\beta_1}]},$$

takes more work.

5.3.4 $I \times 2$ tables

Let $X = 1, 2, \ldots, I$ be an ordinal predictor.

- If the log odds increases linearly with category $X = i$ we have logit $\pi(i) = \alpha + \beta i$.
- If the log risk increases linearly we have $\log \pi(i) = \alpha + \beta i$.
- If the probability increases linearly we have $\pi(i) = \alpha + \beta i$.

If we replace $X = 1, 2, \ldots, I$ by *scores* $u_1 \leq u_2 \leq \cdots \leq u_I$, we get

- logit linear model: logit $\pi(i) = \alpha + \beta u_i$,
- log linear model: $\log \pi(i) = \alpha + \beta u_i$,
- linear model: $\pi(i) = \alpha + \beta u_i$.

- In any of these models testing $H_0 : \beta = 0$ is a test of $X \perp Y$ versus a particular monotone alternative.
- The last of the six is called the Cochran-Armitage linear trend model.
- Tarone and Gart (1980) Showed that the score test (Cochran-Armitage trend test) for a binary linear trend model does not depend on the link function.
- These can all be fit in SAS GENMOD.

```
proc genmod; model present/total = cons / dist=bin link=logit;
proc genmod; model present/total = cons / dist=bin link=log;
proc genmod; model present/total = cons / dist=bin link=identity;
proc genmod; model present/total = score / dist=bin link=logit;
proc genmod; model present/total = score / dist=bin link=log;
proc genmod; model present/total = score / dist=bin link=identity;
```

- The first three use $X = 1, 2, 3, 4, 5$ and the last three use $X = 0.0, 0.5, 1.5, 4.0, 7.0$.
- For this data, the p-values are respectively 0.18, 0.18, 0.28, 0.01, 0.01, 0.13 testing $H_0 : \beta_1 = 0$ using Wald test.
- The Pearson GOF $X^2 = 2.05$ with $p = 0.56$ for the logit model with scores and $X^2 = 5.68$ with $p = 0.13$ for using $1, 2, 3, 4, 5$. The logit model using scores fits better and from this model we reject $H_0 : \beta = 0$ with $p = 0.01$.

Now we have $p-1$ predictors $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{i,p-1})$ and fit

$$Y_i \sim \text{bin}\left(n_i, \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1})}\right).$$

- Many of these predictors may be sets of dummy variables associated with categorical predictors.
- e^{β_j} is now termed the *adjusted* odds ratio. This is how the odds of the event occurring changes when x_j increases by one unit *keeping the remaining predictors constant*.
- This interpretation may not make sense if two predictors are highly related.

An overall test of H_0 : logit $\pi(\mathbf{x}) = \beta_0$ versus H_1 : logit $\pi(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ is generated in PROC LOGISTIC three different ways: LRT, score, and Wald versions. This checks whether some subset of variables in the model is important.

Recall the crab data covariates:

- C = color (1,2,3,4=light medium, medium, dark medium, dark).
- S = spine condition (1,2,3=both good, one worn or broken, both worn or borken).
- W = carapace width (cm).
- Wt = weight (kg).

We'll take $C = 4$ and $S = 3$ as baseline categories.

There are two categorical predictors, C and S, and two continuous predictors W and Wt. Let $Y = 1$ if a randomly drawn crab has one or more satellites and $\mathbf{x} = (C, S, W, Wt)$ be her covariates. An *additive* model including all four covariates would look like

$$\begin{aligned} \text{logit } \pi(\mathbf{x}) &= \beta_0 + \beta_1 I\{C = 1\} + \beta_2 I\{C = 2\} + \beta_3 I\{C = 3\} \\ &\quad + \beta_4 I\{S = 1\} + \beta_5 I\{S = 2\} + \beta_6 W + \beta_7 Wt \end{aligned}$$

This model is fit via

```
proc logistic data=crabs1 descending;
    class color spine / param=ref;
   model y = color spine width weight / lackfit ;
```

The H-L GOF statistic yields $p - value = 0.88$ so there's no evidence of gross lack of fit. The parameter estimates are:

Parameter		DF	Estimate	Standard Error	Wald Chi−Square	Pr > ChiSq
Intercept		1	−9.2734	3.8378	5.8386	0.0157
color	1	1	1.6087	0.9355	2.9567	0.0855
color	2	1	1.5058	0.5667	7.0607	0.0079
color	3	1	1.1198	0.5933	3.5624	0.0591
spine	1	1	−0.4003	0.5027	0.6340	0.4259
spine	2	1	−0.4963	0.6292	0.6222	0.4302
width		1	0.2631	0.1953	1.8152	0.1779
weight		1	0.8258	0.7038	1.3765	0.2407

Color seems to be important. Plugging in $\hat{\beta}$ for β,

$$\begin{aligned} \text{logit } \hat{\pi}(\mathbf{x}) &= -9.27 + 1.61 I\{C=1\} + 1.51 I\{C=2\} + 1.11 I\{C=3\} \\ &\quad -0.40 I\{S=1\} - 0.50 I\{S=2\} + 0.26W + 0.83Wt \end{aligned}$$

Overall checks that one or more predictors are important:

Testing Global Null Hypothesis: BETA=0

Test	Chi−Square	DF	Pr > ChiSq
Likelihood Ratio	40.5565	7	<.0001
Score	36.3068	7	<.0001
Wald	29.4763	7	0.0001

The Type III tests are (1) $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, that color is not needed to explain whether a female has satellite(s), (2) $H_0 : \beta_4 = \beta_5 = 0$, whether spine is needed, (3) $H_0 : \beta_6 = 0$, whether width is needed, and (4) $H_0 : \beta_7 = 0$, whether weight is needed:

Type 3 Analysis of Effects

Effect	DF	Wald Chi−Square	Pr > ChiSq
color	3	7.1610	0.0669
spine	2	1.0105	0.6034
width	1	1.8152	0.1779
weight	1	1.3765	0.2407

The largest p-value is 0.6 for dropping spine condition from the model. When refitting the model without spine condition, we still strongly reject $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$, and the H-L shows no evidence of lack of fit. We have:

Type 3 Analysis of Effects

Effect	DF	Wald Chi−Square	Pr > ChiSq
color	3	6.3143	0.0973
width	1	2.3355	0.1265
weight	1	1.2263	0.2681

We do not reject that we can drop weight from the model, and so we do:

Testing Global Null Hypothesis: BETA=0

Test	Chi−Square	DF	Pr > ChiSq
Likelihood Ratio	38.3015	4	<.0001
Score	34.3384	4	<.0001
Wald	27.6788	4	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi−Square	Pr > ChiSq
color	3	6.6246	0.0849
width	1	19.6573	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi−Square	Pr > ChiSq
Intercept		1	−12.7151	2.7618	21.1965	<.0001
color	1	1	1.3299	0.8525	2.4335	0.1188
color	2	1	1.4023	0.5484	6.5380	0.0106
color	3	1	1.1061	0.5921	3.4901	0.0617
width		1	0.4680	0.1055	19.6573	<.0001

The new model is

$$\text{logit } \pi(\mathbf{x}) = \beta_0 + \beta_1 I\{C = 1\} + \beta_2 I\{C = 2\} \beta_3 I\{C = 3\} + \beta_4 W.$$

We *do not* reject that color can be dropped from the model $H_0 : \beta_1 = \beta_2 = \beta_3$, but we do reject that the dummy for $C = 2$ can be dropped, $H_0 : \beta_2 = 0$. Maybe unnecessary levels in color are clouding its importance.

Let's see what happens when we try to combine levels of C.

```
proc  logistic  data=crabs1 descending;
      class  color  spine  /  param=ref;
      model y = color  width  /  lackfit ;
      contrast  '1  vs  2'  color  1  -1 0;
      contrast  '1  vs  3'  color  1  0  -1;
      contrast  '1  vs  4'  color  1  0  0;
      contrast  '2  vs  3'  color  0  1  -1;
      contrast  '2  vs  4'  color  0  1  0;
      contrast  '3  vs  4'  color  0  0  1;
run;
```

- p-values for combining levels:

Contrast Test Results

Contrast	DF	Wald Chi−Square	Pr > ChiSq
1 vs 2	1	0.0096	0.9220
1 vs 3	1	0.0829	0.7733
1 vs 4	1	2.4335	0.1188
2 vs 3	1	0.5031	0.4781
2 vs 4	1	6.5380	0.0106
3 vs 4	1	3.4901	0.0617

- We reject that we can combine levels $C = 2$ and $C = 4$, and almost reject combining $C = 3$ and $C = 4$. Let's combine $C = 1, 2, 3$ into one category $D = 1$ "not dark" and $C = 4$ is $D = 2$, "dark". See Figure 5.7 (p.188) in next slide.

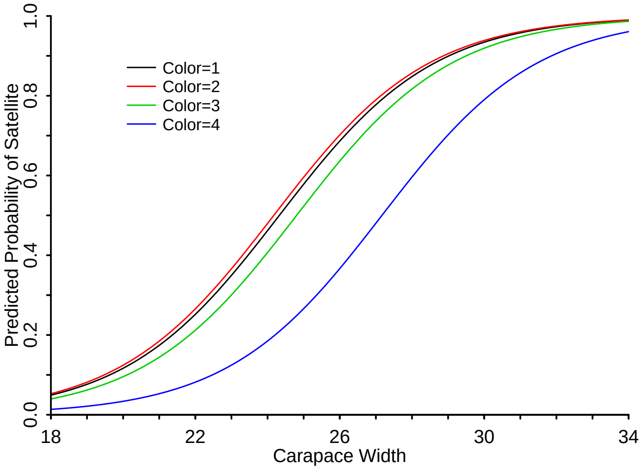

Figure : Predicted probability of satellite presence as a function of width and color

- We include `dark=1; if color=4 then dark=2;` in the DATA step, and fit

```
proc logistic data=crabs1 descending;
 class dark / param=ref ref=first ;
 model y = dark width / lackfit ;
 run;
```

- Annotated output:

Testing Global Null Hypothesis: BETA=0

Test	Chi−Square	DF	Pr > ChiSq
Likelihood Ratio	37.8006	2	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi−Square	Pr > ChiSq
dark	1	6.1162	0.0134
width	1	21.0841	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi−Square	Pr > ChiSq
Intercept		1	−11.6790	2.6925	18.8143	<.0001
dark	2	1	−1.3005	0.5259	6.1162	0.0134
width		1	0.4782	0.1041	21.0841	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
dark 2 vs 1	0.272	0.097	0.764
width	1.613	1.315	1.979

Hosmer and Lemeshow Goodness−of−Fit Test

Chi−Square	DF	Pr > ChiSq
5.5744	8	0.6948

Comments:

- The odds of having satellite(s) significantly decreases by a little less than a third, 0.27, for dark crabs regardless of width.
- The odds of having satellite(s) significantly increases by a factor of 1.6 for every *cm* increase in carapice width regardless of color.
- Lighter, wider crabs tend to have satellite(s) more often.
- The H-L GOF test shows no gross LOF.
- We didn't check for interactions. If an interaction between color and width existed, then the odds ratio of satellite(s) for dark versus not dark crabs would change with how wide she is.

Interactions and quadratic effects

- An additive model is easily interpreted because an odds ratio from changing values of one predictor does not change with levels of another predictor. However, often this incorrect and we may introduce additional terms into the model such as interactions.
- An interaction between two predictors allows the odds ratio for increasing one predictor to change with levels of another. For example, in the last model fit the odds of having satellite(s) decreases by 0.27 for dark crabs vs. not dark *regardless of carapace width*.
- A two-way interaction is defined by multiplying the variables together; if one or both variables are categorical then all possible pairings of dummy variables are considered.

Example: Say we have two categorical predictors, $X = 1, 2, 3$ and $Z = 1, 2, 3, 4$. An additive model is

$$\begin{aligned} \text{logit } \pi(X, Z) &= \beta_0 + \beta_1 I\{X = 1\} + \beta_2 I\{X = 2\} \\ &\quad + \beta_3 I\{Z = 1\} + \beta_4 I\{Z = 2\} + \beta_5 I\{Z = 3\}. \end{aligned}$$

The model that includes an interaction between X and Z adds $(3 - 1)(4 - 1) = 6$ additional dummy variables accounting for all possible ways, i.e. all levels of Z, the log odds can change between from $X = i$ to $X = j$. The new model is rather cumbersome:

$$\begin{aligned} \text{logit } \pi(X, Z) &= \beta_0 + \beta_1 I\{X = 1\} + \beta_2 I\{X = 2\} \\ &\quad + \beta_3 I\{Z = 1\} + \beta_4 I\{Z = 2\} + \beta_5 I\{Z = 3\} \\ &\quad + \beta_6 I\{X = 1\} I\{Z = 1\} + \beta_7 I\{X = 1\} I\{Z = 2\} \\ &\quad + \beta_8 I\{X = 1\} I\{Z = 3\} + \beta_9 I\{X = 2\} I\{Z = 1\} \\ &\quad + \beta_{10} I\{X = 2\} I\{Z = 2\} + \beta_{11} I\{X = 2\} I\{Z = 3\}. \end{aligned}$$

- In PROC GENMOD and PROC LOGISTIC, categorical variables are defined through the CLASS statement and all dummy variables are created and handled internally.
- The Type III table provides a test that the interaction can be dropped; the table of regression coefficients tell you whether individual dummies can be dropped.
- Let's consider the crab data again, but consider an interaction between categorical D and continuous W:

```
proc logistic data=crabs1 descending;
 class dark / param=ref ref=first ;
 model y = dark width dark*width / lackfit ;
```

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
dark	1	0.9039	0.3417
width	1	20.7562	<.0001
width*dark	1	1.2686	0.2600

We accept that the interaction is not needed.

Let's consider the interaction model anyway, for illustration:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi−Square	Pr > ChiSq
Intercept		1	−12.8116	2.9577	18.7629	<.0001
dark	2	1	6.9578	7.3182	0.9039	0.3417
width		1	0.5222	0.1146	20.7562	<.0001
width∗dark	2	1	−0.3217	0.2857	1.2686	0.2600

The model is:

$$\text{logit } \pi(D, W) = -12.81 + 6.96 I\{D = 2\} + 0.52W - 0.32 I\{D = 2\}W.$$

The odds ratio for the probability of satellite(s) going from $D = 2$ to $D = 1$ is estimated

$$\frac{P(Y = 1|D = 2, W)/P(Y = 0|D = 2, W)}{P(Y = 1|D = 1, W)/P(Y = 0|D = 1, W)} = \frac{e^{-12.81+6.96+0.52W-0.32W}}{e^{-12.81+0.52W}} = e^{6.96-0.32W}.$$

How about the odds ratio going from W to $W + 1$?

- For a categorical predictor X with I levels, adding $I-1$ dummy variables allows for a different event probability at each level of X.
- For a continuous predictor Z, the model assumes that the log-odds of the event increases *linearly* with Z. This may or may not be a reasonable assumption, but can be checked by adding nonlinear terms, the simplest being Z^2.
- Consider a simple model with continuous Z:

$$\text{logit } \pi(Z) = \beta_0 + \beta_1 Z.$$

LOF from this model can manifest itself in rejecting a GOF test (Pearson, deviance, or H-L) or a residual plot that shows curvature.

Adding a quadratic term

$$\text{logit } \pi(Z) = \beta_0 + \beta_1 Z + \beta_2 Z^2,$$

may improve fit and allows testing the adequacy of the simpler model via $H_0 : \beta_2 = 0$. Higher order powers can be added, but the model can become unstable with, say, higher than cubic powers. A better approach might be to fit a *generalized additive model* (GAM):

$$\text{logit } \pi(Z) = f(Z),$$

where $f(\cdot)$ is estimated from the data, often using splines.
However, we will not discuss this in this course!
Adding a simple quadratic term can be done, e.g.,

```
proc logistic; model y/n = z z*z;
```

Should you always toss in a dispersion term ϕ?

Here's some SAS code for a made-up data:

```
data example;
 input x y n @@; x_sq=x*x;
 datalines;
 -2.0 86 100 -1.5 58 100 -1.0 25 100 -0.5 17 100 0.0 10 100
  0.5 17 100  1.0 25 100
 ;
proc genmod; * fit simple linear term in x & check for overdispersion;
    model y/n = x / link=logit dist=bin;
proc genmod; * adjust for apparent overdispersion;
    model y/n = x / link=logit dist=bin scale=pearson;
proc genmod; * what if instead we try a more flexible mean?;
    model y/n = x x_sq / link=logit dist=binom;
proc logistic; * residual plots from simpler model;
    model y/n = x; output out=diag1 reschi=p h=h xbeta=eta;
data diag2; set diag1; r=p/sqrt(1-h);
proc gplot; plot r*x; plot r*eta; run;
```

Output from fit of logistic model with logit link:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	5	74.6045	14.9209
Pearson Chi−Square	5	79.5309	15.9062

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi−Square	Pr > ChiSq
Intercept	1	−1.3365	0.1182	−1.5682	−1.1047	127.77	<.0001
x	1	−1.0258	0.0987	−1.2192	−0.8323	108.03	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

The coefficient for x is highly significant. Note that $P(\chi^2_5 > 74.6) < 0.0001$ and $P(\chi^2_5 > 79.5) < 0.0001$. Evidence of overdispersion? There's good replication here, so certainly *something* is not right with the model.

Let's include a dispersion parameter ϕ:

```
                 Criteria For Assessing Goodness Of Fit

Criterion                   DF          Value        Value/DF
Deviance                     5        74.6045         14.9209
Scaled Deviance              5         4.6903          0.9381
Pearson Chi-Square           5        79.5309         15.9062
Scaled Pearson X2            5         5.0000          1.0000

                  Analysis Of Parameter Estimates

                          Standard     Wald 95% Confidence      Chi-
Parameter  DF  Estimate     Error            Limits            Square   Pr > ChiSq
Intercept   1   -1.3365    0.4715     -2.2607     -0.4123       8.03       0.0046
x           1   -1.0258    0.3936     -1.7972     -0.2543       6.79       0.0092
Scale       0    3.9883    0.0000      3.9883      3.9883
```

We have $\hat{\phi} = 3.99$ and the standard errors are increased by this factor.
The coefficient for x is still significant.
Problem solved!!! Or is it?

Instead of adding ϕ to a model with a linear term, what happens if we allow the mean to be a bit more flexible?

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	4	1.7098	0.4274
Pearson Chi-Square	4	1.6931	0.4233

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	−1.9607	0.1460	−2.2468	−1.6745	180.33	<.0001
x	1	−0.0436	0.1352	−0.3085	0.2214	0.10	0.7473
x_sq	1	0.9409	0.1154	0.7146	1.1671	66.44	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Here, we are *not* including a dispersion term ϕ. There is no evidence of overdispersion when the *mean is modeled correctly*. Adjusting SE's using the quasilikelihood approach relies on *correctly modeling the mean*, otherwise ϕ becomes a measure of dispersion of data about *an incorrect mean*. That is, ϕ attempts to pick up the slop left over from specifying a mean that is too simple.

A correctly specified mean can obviate overdispersion. How to check if the mean is okay? Hint:

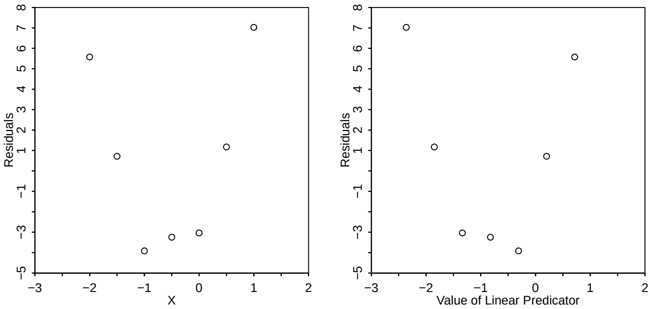

Figure : Residual plots r_i versus $(X_i\ \eta_i)$ for made-up data.

5.4.8 Estimating an Average Causal Effect

- In many applications the explanatory variable of primary interest specifies two groups to be compared while adjusting for the other explanatory variables in the model.
- Let $X_1 = 0, 1$ denote this two groups.
- As an alternative effect summary to the log odds ratio $\hat{\beta}_1$, the estimated average causal effect is

$$\frac{1}{n}\sum_i [\hat{\pi}(\mathbf{x}_{i1} = 1, x_{i2}, ..., x_{ip}) - \hat{\pi}(\mathbf{x}_{i1} = 0, x_{i2}, ..., x_{ip})]$$

- Estimating an average causal effect is natural for experimental studies, and received much attention for non-randomized studies.

5.5 Fitting logistic regression models

The data are $(\mathbf{x}_i, Y_i)$ for $i = 1, \ldots, N$.

The model is

$$Y_i \sim \text{bin}\left(n_i, \frac{e^{\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\boldsymbol{\beta}'\mathbf{x}_i}}\right).$$

The pmf of Y_i in terms of $\boldsymbol{\beta}$ is

$$p(y_i;\boldsymbol{\beta}) = \binom{n_i}{y_i}\left[\frac{e^{\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\boldsymbol{\beta}'\mathbf{x}_i}}\right]^{y_i}\left[1-\frac{e^{\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\boldsymbol{\beta}'\mathbf{x}_i}}\right]^{n_i-y_i}.$$

The likelihood is the product of all N of these and the log-likelihood simplifies to

$$L(\boldsymbol{\beta}) = \sum_{j=1}^{p}\beta_j\sum_{i=1}^{N}y_i x_{ij} - \sum_{i=1}^{N}\log\left[1+\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)\right] + \text{constant}.$$

The likelihood (or score) equations are obtained by taking partial derivatives of $L(\boldsymbol{\beta})$ with respect to elements of $\boldsymbol{\beta}$ and setting equal to zero. Newton-Raphson is used to get $\hat{\beta}$, see 5.5.4 if interested.
The inverse of the covariance of $\hat{\boldsymbol{\beta}}$ has ij^{th} element

$$-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_i\partial\beta_j} = \sum_{s=1}^{N} x_{si}x_{sj}n_s\pi_s(1-\pi_s),$$

where $\pi_s = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_s}}{1+e^{\boldsymbol{\beta}'\mathbf{x}_s}}$. The *estimated* covariance matrix $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})$ is obtained by replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$. This can be rewritten

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = \{\mathbf{X}'\text{diag}[n_i\hat{\pi}_i(1-\hat{\pi}_i)]\mathbf{X}\}^{-1}.$$

Existence of finite $\hat{\beta}$

- Estimates $\hat{\beta}$ exist except when data are perfectly separated.
- Complete separation happens when a linear combination of predictors perfectly predicts the outcome. Here, there are an infinite number of perfect fitting curves that have $\hat{\beta} = \infty$. Essentially, there is a value of x that perfectly separates the 0's and 1's. In two-dimensions there would be a line separating the 0's and 1's.
- Quasi-complete separation happens when there's a line that separates 0's and 1's but there's some 0's and 1's on the line. We'll look at some pictures.
- The end result is that the model will appear to fit but the standard errors will be absurdly large. This is the *opposite* of what's really happening, that the data can be perfectly predicted.