Chapter 14

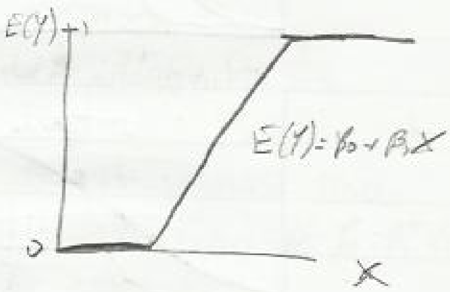
Regression Models w/ Binary Response Variable

- When response variable only has 2 qualitative outcomes

$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad Y_i = 0, 1$

$E(Y_i) = \beta_0 + \beta_1 X_i$

Y_i	Prb
1	$P(Y_i = 1) = \pi_i$
0	$P(Y_i = 0) = 1 - \pi_i$



$E(Y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i = P(Y_i = 1)$

$E(Y_i) = \beta_0 + \beta X_i = \pi_i$

Special Problems

1) When $Y_i = 1 \quad \epsilon_i = 1 - \beta_0 - \beta_1 X$ * Non normal Error Terms

$Y_i = 0 \quad \epsilon_i = -\beta_0 - \beta_1 X_i$

2) Non-constant variance

$\sigma^2\{Y_i\} = E\{(Y_i - E\{Y_i\})^2\} = (1-\pi_i)^2 \pi_i + (0 - \pi_i)^2 (1 - \pi_i)$

or

$\sigma^2\{Y_i\} = \pi_i(1 - \pi_i) = (E\{Y_i\})(1 - E\{Y_i\})$

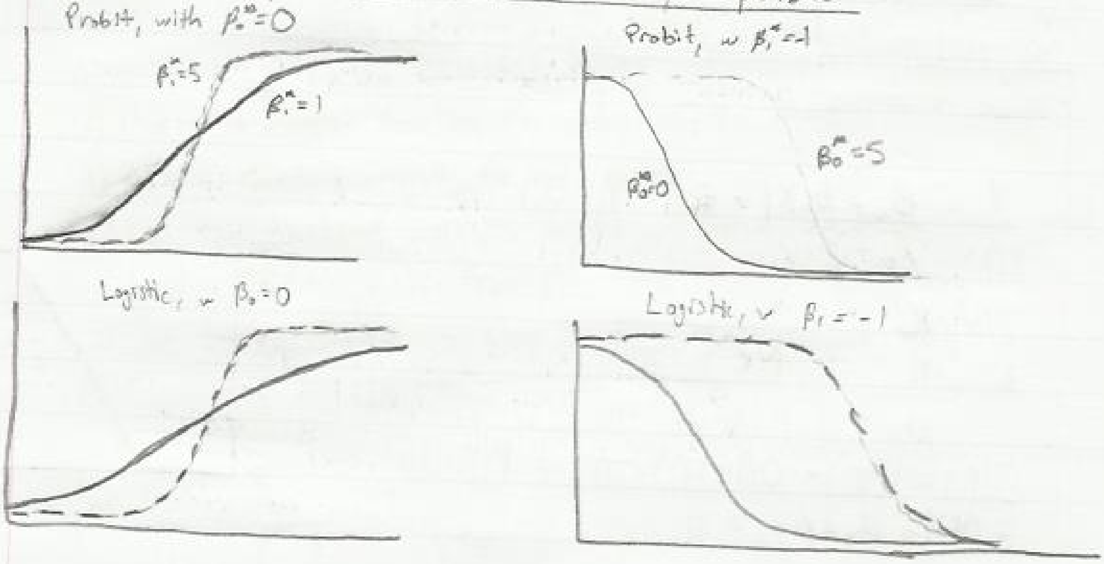
$= (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i)$

* Depends on x thus non constant

3) Constraints on Responce function

$0 \leq E\{Y_i\} = \pi \leq 1$

Sigmoidal Response Functions for Binary Responses



Probit $\quad Y_i^c = \beta_0^c + \beta_1^c X_i + \epsilon_i^c$

$$\ni Y_i = \begin{cases} 1 & \text{if, say } Y_i^c \le 38 \\ 0 & Y_i^c > 38 \end{cases}$$

$$\begin{aligned} P(Y_i=1) = \pi_i &= P(Y_i^c \le 38) \\ &= P(\beta_0^c + \beta_1^c X_i + \epsilon_i^c \le 38) \\ &= P(\epsilon_i^c \le 38 - \beta_0^c - \beta_1^c X_i) \\ &= P\left(\frac{\epsilon_i^c}{\sigma_c} \le \frac{38-\beta_0^c}{\sigma_c} - \frac{\beta_1^c}{\sigma_c} X_i\right) \\ &= P(Z \le \beta_0^* + \beta_1^* X_i) \end{aligned}$$

$\ni \beta_0^* = (38-\beta_0^c)/\sigma_c$, $\beta_1^* = -\beta_1^c/\sigma_c$ $\ni Z = \epsilon_i^c/\sigma_c \ni Z \sim \text{std } N$

$\Rightarrow P(Y_i=1) = \Phi(\beta_0^* + \beta_1^* X_i) \quad \ni P(Z \le z) = \Phi(z)$

Thus $E(Y_i) = \pi = \Phi(\beta_0^* + \beta_1^* X_i)$

& $\Phi^{-1}(\pi_i) = \pi_i' = \beta_0^* + \beta_1^* X_i$

And finally, $\pi' = \beta_0^* + \beta_1^* X_i$ is the <u>Probit Responce Function</u>

or <u>linear predictor</u>

Charactoristics of π' for probit & logistic

1) Bounded between 0 and 1 and reaches those limits asymptotically
2) As β_1^* increases (for $\beta_1^* > 0$) the mean function becomes more switch like
3) Changing the sign of β_1^* from positive to negative we create a monotone decreasing response
4) Increasing or decreasing the intercept β_0^* shifts the mean response function horizontal
5) Symmetry property If we let $Y_i' = 1 - Y_i$ then it follows since $\phi(z) = 1 - \Phi(z) \rightarrow Y(Y_i' = 1) = P(Y_i = 0) 1 - \Phi(\beta_0^* + \beta_1^* X_i) = \phi(-\beta_0^* - \beta_1^* X_i)$

* Max likelihood estimater are parameter estimater

$\boxed{\text{SAS}}$
```
proc genmod;
   model y= x1 x2 x3 x4 / dist=poisson link=probit;
run;
```

Logistic Mean Response Function

- Logistic is close to standard normal but it has fatter tails
- $F_L^{-1}(\pi_i) = \beta_0 + \beta_1 X_i = \pi_i' \Leftarrow$ logit response function
  $= \log_e\left(\frac{\pi_i}{1-\pi_i}\right)$ $\quad \frac{\pi_i}{1-\pi_i} \leftarrow$ the odds

$\boxed{\text{SAS}}$
```
proc logistic;  *Or we can use genmod w/ link=logit.
model y(event='1')= x1 x2 x3 / selection= forward  slstay=.1  ;
                                          backward slentry=.1
run;
```

Complementary Log-Log Response Function - skewed right (extreme value / Gumbel)

$\pi_i = 1 - \exp(-\exp(\beta_0^G + \beta_1^G X_i))$ mean response function

$\pi_i' = \log(-\log(1-\pi(X_i))) = \beta_0^G + \beta_1^G X_i$

$\boxed{\text{SAS}}$
```
proc logistic;
   model y(event='1')= x1 x2 x3 / link= cloglog;
run;
```

14.3 Model Interpretation

$$Y_i \sim \text{Bernoulli}\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)$$

Simple Logistic Model: $Y_i = E(Y_i) + \epsilon_i$

$$E(Y_i) = \pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

How odds ratio changes when x increases by 1: $\dfrac{\pi(x+1)/[1-\pi(x+1)]}{\pi(x)/[1-\pi(x)]} = \dfrac{e^{\beta_0+\beta_1 x+\beta_1}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$

interpretation: $\forall$ 1 unit increase in x the odds of an event occurring increases by a factor of e^{β_1}

14.4 Multiple Logistic Regression

$$\underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \quad \underline{X}_{p\times 1} = \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_{p-1} \end{bmatrix} \quad X_{i\,p\times 1} = \begin{bmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{i,p-1} \end{bmatrix}$$

Then,

$$\underline{X}'\underline{\beta} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

$$\underline{X}_i'\underline{\beta} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$$

$$E\{Y\} = \frac{\exp(\underline{X}'\underline{\beta})}{1+\exp(\underline{X}'\underline{\beta})} = [1 + \exp(-\underline{X}'\underline{\beta})]^{-1}$$

$$\pi' = \log_e\left(\frac{\pi}{1-\pi}\right) = \underline{X}'\underline{\beta}$$

$$\pi' = \underline{X}'\underline{\beta}$$

Thus, for Y_i independent Bernoulli random variables

$$E(Y_i) = \pi_i = \frac{\exp(\underline{X}_i'\underline{\beta})}{1+\exp(\underline{X}_i'\underline{\beta})}$$

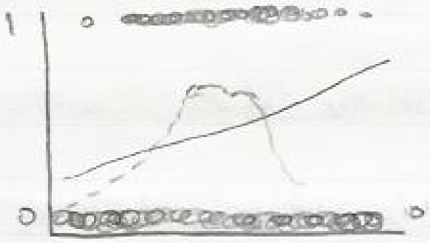
Maximum Likelihood Estimators $\underline{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$

$$\hat{\pi} = \frac{\exp(\underline{X}'\underline{b})}{1+\exp(\underline{X}'\underline{b})} = (1+\exp(-\underline{X}'\underline{b}))^{-1}$$

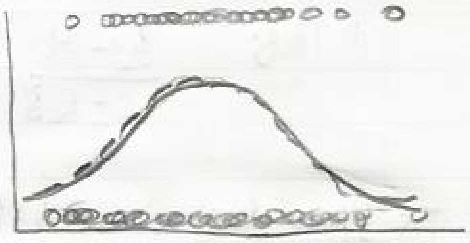
$$\hat{\pi}_i = \frac{\exp(\underline{X}_i'\underline{b})}{1+\exp(\underline{X}'\underline{b})} = (1+\exp(-\underline{X}_i'\underline{b}))^{-1}$$

Polynomial Logistic Regression

$$\pi'(x) = \beta_0 + \beta_{11}x + \beta_{22}x^2 + \cdots + \beta_{kk}x^k$$
where $x = X - \overline{X}$



First Order



Second order

*SAS is the same except we change the model statement to a second order model

14.5 Inferences about Regression Parameters

* For large sample sizes, under generally applicable conditions, maximum likelihood estimators for logistic regression are approximately normally distributed

· Let G denote the matrix of second-order partial derivatives of the loglikelihood function

$$\log_e L(\beta) = \sum_i^n Y_i(X_i'\beta) - \sum_i^n \log_e[1+e^{X_i'\beta}]$$

w/ regard to the parameters $\beta_0, \beta_1, \dots \beta_{p-1}$

$G_{p\times p} = [g_{ij}]$ $i=0,1,\dots p-1$ $j=0,1,\dots p-1$ [Hessian Matrix]

$\ni$

$$G_{00} = \frac{\partial^2 \log_e L(\beta)}{\partial \beta_0^2}$$

$$G_{01} = \frac{\partial^2 \log_e L(\beta)}{\partial \beta_0 \partial \beta_1}$$

$\vdots$

Note $s^2\{b\} = ([-g_{ij}]_{\beta=b})^{-1}$

Thus $\dfrac{b_k - \beta_k}{s\{b_k\}} \sim Z$

<u>Wald test</u> $H_0: \beta_k = 0$ | $H_0: \beta_1 \le 0$

$H_a: \beta_k \ne 0$ | $H_a: \beta_1 > 0$

$z^* = \frac{b_k}{s\{b_k\}}$ | $z^* = \frac{b_1}{s\{b_1\}}$

$|z^*| > z(1-\alpha/2)$ reject | $z^* > z(1-\alpha/2)$ reject

[SAS] this is given by proc logistic as part of the default output table 'Analysis of Maximum Likelihood Estimates"

<u>Interval Estimation</u>

$1-\alpha$ confidence limits for β_k: $b_k \pm z(1-\alpha/2)s\{b_k\}$

[SAS] Add '/cl' to your model statement in proc logistic

Test whether Several $\beta_k = 0$: Likelihood Ratio Test

Full model: $\pi = [1 + \exp(-\underline{X}'\underline{\beta}_F)]^{-1}$

$\ni X'\beta_F = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$

evaluate $L(\beta)$ when $\beta_F = b_F \Rightarrow L(F)$

Want to test: $H_0: \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$

H_a: at least one above $\neq 0$

Reduced model: $\pi = [1 + \exp(-\underline{X}'\underline{\beta}_R)]^{-1}$

$\ni X'\beta_R = \beta_0 + \beta_1 X_1 + \cdots \beta_{q-1} X_{q-1}$

evaluate $L(\beta)$ when $\underline{\beta}_R = \underline{b}_R \Rightarrow L(R)$

$G^2 = -2 \log_e \left[\frac{L(R)}{L(F)}\right] = -2[\log_e L(R) - \log_e L(F)]$

When $G^2 > \chi^2(1-\alpha, p-q)$ reject

Automatic Model Selection

Stepwise, forward and backward are the same as pg. 40.

SAS

```
proc logistic;
  model y(event='1') = x1 x2 x2 x2*x2 / selection  { stepwise;
                                                     forward slentry=.1;
                                                     backward slstay=.1;
                         'best' subsets  ->          score;
```

Pearson Chi-Square Goodness of Fit

$H_0: E\{Y\} = [1 + \exp(-\underline{X}'\underline{\beta})]^{-1}$

$H_a: E\{Y\} \neq [1 + \exp(-\underline{X}'\underline{\beta})]^{-1}$

$$X^2 = \sum_{j=1}^{c} \frac{(Y_{.j} - n_j \hat{\pi}_j)^2}{n_j \hat{\pi}_j (1 - \hat{\pi}_j)}$$

If $X^2 > \chi^2(1-\alpha, c-p)$ Reject Ha

Deviance Goodness of fit test

Reduced Model: $E\{Y_{ij}\} = [1 + \exp(-\underline{X}_j' \underline{\beta})]^{-1}$

Full Model: $E\{Y_{ij}\} = \pi_j$ $\quad j = 1, 2, \dots c$ Full model

let $p_j = \frac{Y_{\cdot j}}{n_j}$ $\quad j = 1, 2, \dots c$

$\hat{\pi}_j$ = reduced model estimate of π_j

$$G^2 = -2[\log_e(L(R)) - \log_e(L(F))]$$
$$= -2 \sum_{i}^{c} \left[Y_{\cdot j} \log_e\left(\frac{\hat{\pi}_j}{p_j}\right) + (n_j - Y_{\cdot j}) \log_e\left(\frac{1-\hat{\pi}_j}{1-p_j}\right)\right]$$
$$= DEV(X_0, X_1, \dots, X_{p-1})$$

$H_0: E\{Y\} = [1 + \exp(-\underline{X}'\underline{\beta})]^{-1}$

$H_a: E\{Y\} \neq [1 + \exp(-\underline{X}'\underline{\beta})]^{-1}$

If $DEV(X_0, X_1, \dots X_{p-1}) > \chi^2(1-\alpha, c-p)$ reject

SAS

```
proc logistic;
  model y(event='1') = x1 x2 x3 /aggregate scale=none;
run;
```

output

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance				P values
Pearson				

Hosmer-Lemeshow Goodness of fit test

* Assumes about the same number of cases in each class
  - can reform groups according to fitted logit values

SAS

```
proc logistic;
 model y = x1 x2 x3 / lackfit;
run;
```

output

Hosmer and Lemeshow Goodness-of fit test

Chi-sq	DF	Pr > Chisq
		pvalue

14.8 Logistic Regression Diagnostics

$$e_i = \begin{cases} 1-\hat{\pi}_i & \text{if } Y_i = 1 \\ -\hat{\pi}_i & \text{if } Y_i = 0 \end{cases}$$

Pearson Residuals $\quad r_{p_i} = \dfrac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$

Studentized Pearson Residual $\quad r_{sp_i} = \dfrac{r_p}{\sqrt{1-h_{ii}}}$

h_{ii} is the i^{th} diagonal of $\underline{H} = \underline{\hat{W}}^{1/2}\underline{X}(\underline{X}'\underline{\hat{W}}\,\underline{X})^{-1}\underline{X}'\underline{\hat{W}}^{\frac{1}{2}}$

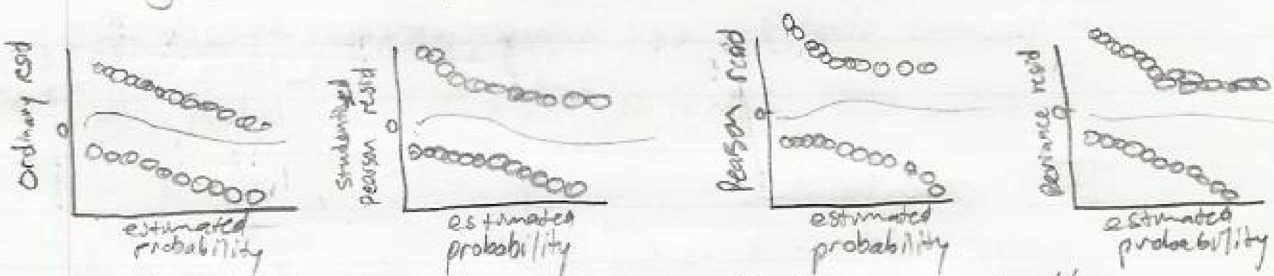
Deviance Residual $\quad dev_i = \text{sign}(Y_i - \hat{\pi}_i)\sqrt{-2[Y_i \log_e(\hat{\pi}_i) + (1-Y_i)\log_e(1-\hat{\pi})]}$

SAS

```
proc logistic;
model y(event='1') = x1 x2 x3 / influence lackfit;
run;
```

* this will give you everything but rsp, you will have to calculate this using an output statement

Diagnostic Residual plots



* Should have a loess curve that is approximatley horizontal at 0, any departure suggests our model is inadequate

$\boxed{\text{SAS}}$
```
proc logistic;
model y(event='1') = x1 x2;
output=out p=p h=h reschi=rp resdev=r;
run
proc sgscatter data=out;
    plot r*(p x1 x2)/loess;
run;
```

Detection of influential points

$\Delta X_i^2 = X^2 - X_{(i)}^2$ = change in pearson goodness of fit statistic when we remove the i^{th} observation

$\approx r_{sp_i}^2$

$\Delta dev_i = DEV - DEV_{(i)}$ = change in deviance goodness of fit statistic when we remove the i^{th} observation

$\approx h_{ii} r_{sp_i}^2 - dev_i^2$

The larger ΔX_i^2 and Δdev_i the more influential

Cooks' Distance Measures the standardized change in the fitted response vector $\hat{Y}$ when the i^{th} case is deleted

$$D_i = \frac{r_{P_i}^2 h_{ii}}{p(1-h_{ii})^2}$$

14.9 Inferences about Mean response

Point Estimation

$$\underline{X}_n = \begin{bmatrix} 1 \\ X_{n,1} \\ \vdots \\ X_{n,p-1} \end{bmatrix}$$ is to be predicted

$$\pi_n = [1 + \exp(-\underline{X}'\underline{\beta})]^{-1}$$

$$\hat{\pi}_n = [1 + \exp(-\underline{X}'\underline{b})]^{-1}$$

Interval Estimation

$$\hat{\pi}_n' \pm z(1-\alpha/2)\, s\{\hat{\pi}_n'\} \quad \text{for } \pi_n'$$

$$U = \hat{\pi}_n' + z(1-\alpha/2)\, s\{\hat{\pi}_n'\} \Rightarrow U^* = [1 + \exp(-U)]^{-1}$$
$$L = \hat{\pi}_n - z(1-\alpha/2)\, s\{\hat{\pi}_n'\} \quad L^* = [1 + \exp(-L)]^{-1}$$

for mean resp. of π_n

Skip 14.10 - 14.12

<u>14.13 Poisson Regression</u> When response variable is a count but a large count is rare

<u>Poisson Dist</u>

$$f(Y) = \frac{\mu^Y e^{-\mu}}{Y!} \quad \ni Y = 0, 1, 2, \ldots$$

$E\{Y\} = \mu$

$\sigma^2(Y) = \mu$

Note: when count response pertains to different units of time we have $f(Y) = \frac{(t\mu)^Y e^{-t\mu}}{Y!}$ <u>Note</u> $\log(t_i)$ is called an offset

<u>Regression Model</u> $Y_i = E\{Y_i\} + \epsilon_i \quad i = 1, 2, \ldots n$

- we use $\mu(\underline{X}_i, \underline{\beta})$ to denote the function that relates the mean response, μ_i, to $\underline{X}_i$

  ie:
  - $\mu_i = \mu(\underline{X}_i, \underline{\beta}) = \underline{X}_i' \underline{\beta}$
  - * $\mu_i = \mu(\underline{X}_i, \underline{\beta}) = \exp(\underline{X}_i' \underline{\beta})$ * Most common
  - $\mu_i = \mu(\underline{X}_i, \underline{\beta}) = \log_e(\underline{X}_i' \underline{\beta})$

  $\ni \mu_i$ is non negative

log-link $\rightarrow \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i\,p-1}$

identity-link $\rightarrow \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots \beta_{p-1} x_{i\,p-1}$

<u>interpretation of log link:</u> increasing x_j by one, all else constant, increases the mean response by factor e^{β_j}

$\boxed{\text{SAS}}$ proc genmod; * $Sa_i \sim Pois(e^{\alpha + \beta W_i})$
model y = x / dist=poi link=log;
run;
proc genmod; * $Sa_i \sim PoB(\alpha + \beta W_i)$
model y=x / dist = poi link=identity;
run;
proc genmod; * $Sa_i \sim Pois(e^{\alpha + \beta_1 W_i + \beta_2 W_i^2})$
model Y=x xsq / dist= poi link= log;
run;

for offset

proc genmod;
model y= x / dist=poi link=log offset = logtime;
run;

* Follows similar diagnostics as logit regression

14.14 Generalized linear models

1) $Y_1, \dots Y_n$ ar n independent responses that follow a probability distribution belonging to the exponential family ∋ $E\{Y_i\} = \mu_i$
2) A linear predictor based on $X_{i1}, X_{i2}, \dots X_{i,p-1}$
$$\underline{X}_i' \underline{\beta} = \beta_0 + \beta_1 X_{i1} + \dots \beta_{p-1} X_{i,p-1}$$
3) The link function relates the linear predictor to the mean resp
$$\underline{X}_i' \underline{\beta} = g(\mu_i)$$
4) May have nonconstant variance but σ^2 must be a function of the predictors