In case of linear models the true dist^{n} f is of the form $g(y;\theta)$ for some $\underline{\theta}$ if $f \underline{M} \in \mathcal{C}(X)$.

$\Leftrightarrow \lambda = 0$

When this holds $\frac{n\hat{\sigma}^2}{\sigma_0^2} \sim \chi^2_{n-p}$

For a χ^2_p r.v. X, $E\left(\frac{1}{X}\right) = \frac{1}{p-2}$

So, $E\left[\frac{\sigma_0^2}{\hat{\sigma}^2}\right] = \frac{n}{n-p-2}$

& $E[2\Delta] = n\,E\left[\log(2\pi\hat{\sigma})^2\right] + \frac{n(n+p)}{n-p-2}$

- Both 2Δ and AIC have expectationally of order $O(n)$.
- The bias in estimating $E[2\Delta]$ by AIC is

$$E[AIC - 2\Delta] = n + 2(p+1) - \frac{n(n+p)}{n-p-2}$$

$$= 2(p+1) - \frac{2n(p+1)}{n-p-2} = O(1)$$

- Although the bias is of smaller order than $E[2\Delta]$, we can get an exactly unbiased of $E[2\Delta]$ by using a modified criterion.

$$AIC_c = n\log 2\pi\hat{\sigma}^2 + \frac{n(n+p)}{n-p-2}.$$

In many cases, AIC_c is a much better estimate of 2Δ than AIC.

If μ is not in $\mathcal{E}(x)$, then

$$n E[\hat{\sigma}^2] = \sigma_0^2 (n-p+\lambda)$$

$$E[RSS_p] = E(ME) + \sigma^2(n-2p)$$

$$= \lambda\sigma^2 + \sigma^2 p + \sigma^2(n-2p)$$

$$E\left[\frac{\sigma_0^2}{\hat{\sigma}^2}\right] \approx \frac{\sigma_0^2}{E[\hat{\sigma}^2]} = \frac{n}{\lambda+n-p}$$

So, $$E[AIC - 2\Delta] \approx n + 2(p+1) - \frac{(n+p+\lambda)\,n}{\lambda+n-p}$$

$$= 2\left[p+1 - \frac{np}{\lambda+n-p}\right]$$

$$\approx 2$$

$\frac{np}{\lambda+n-p} \approx p$ for large n.

- A slightly different criterions is if we assume σ^2 is unknown.
- So the no of unknown is p and

$$-2\log\left(g(\underline{y}; \hat{\theta}(\underline{y})\right) = n\log(2\pi\sigma^2) + \frac{(\underline{y}-X\hat{\underline{\beta}})'(\underline{y}-X\hat{\underline{\beta}})}{\sigma^2}$$

$$AIC = \frac{RSS_p}{\sigma^2} + 2p$$

$\longrightarrow$ very similar to C_p.

- A obvious generation of this version of AIC is to consider the criterion of the form.

$$AIC = \frac{RSS_p}{\sigma^2} + a_n p$$

- a_n is allowed to depend on n
- $a_n = \log n \longrightarrow$ Bayes information criterion. (BIC).

Polynomial Regression:

Polynomial in one variable

Polynomial in one variable:

· A major problem when fitting a high degree polynomial.

· Set $x_{ij} = x_i^j$, $K = p-1$ $(\leq n-1)$ in general multiple regression set up.

$$y_i = \beta_0 x_i^0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_K x_i^K, \quad i = 1, 2, \ldots, n$$

$x_0 = 1$

· When $K > 6$, X become

Assume that $x_i \overset{appr}{\sim} U[0,1]$, then for large n

$$(X'X)_{(r,s)} = n \frac{1}{n} \sum_{i=1}^{n} x_i^r x_i^s = n \frac{1}{n} \left[\sum_{i=0}^{n} x_i^r x_i^s \right]$$

$$\approx n \int_0^1 x^r x^s \, dx$$

$$= n \int_0^1 x^{r+s} \, dx = \frac{n}{r+s+1}$$

So, $(X'X) \longrightarrow (K+1) \times (K+1)$ principle minor of the Hilbert matrix.

$$H = \begin{pmatrix} 1 & 1/2 & 1/3 & 1/4 & \cdots \\ 1/2 & 1/3 & 1/4 & 1/5 & \cdots \\ 1/3 & 1/4 & 1/5 & 1/6 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix}$$

When $k=9$, the inverse of H_{10} (10×10 principle minor of H) has elements of the ~~neg~~ magnitude 3×10^{10}.

If a small error of 10^{-10} in one element of $X'\underline{y}$

$\longrightarrow$ will lead to an error of ~~deviation~~ about 3 in an element of $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$

- Two things can be done.

(1) Normalize x_i so that they run from -1 to 1.

$$x_i' = \frac{2x_i - \max(x_i) - \min(x_0)}{\max(x_i) - \min(x_i)}$$

(2) Use orthogonal polynomial.

Consider the model.

$$y_i = \gamma_0 \phi_0(x_i) + \gamma_1 \phi_1(x_i) + \cdots + \gamma_k \phi_k(x_i) + \epsilon_i$$

When $\phi_r(x_i) \longrightarrow r$th degree polynomial in x_i, $r=0,1,\ldots,k$.

$$\sum_{i=1}^{n} \phi_r(x_i)\,\phi_s(x_i) = 0 \quad \forall\, r, s, \quad r \neq s.$$

Then $\underline{y} = X\underline{\gamma} + \underline{\epsilon}$

$$X = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_k(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \cdots & \phi_k(x_2) \\ \vdots & & & \\ \phi_0(x_n) & \cdots & & \phi_k(x_n) \end{pmatrix}$$

$$X'X = \begin{bmatrix} \sum \phi_0^2(x_i) & 0 & \cdots & 0 \\ & \sum \phi_1^2(x_i) & & \\ 0 & & \ddots & \\ & & & \sum_i \phi_k^2(x_i) \end{bmatrix}$$

$$\hat{\underline{\gamma}} = (X'X)^{-1}X'\underline{y} \Rightarrow \hat{\gamma}_r = \frac{\sum_i \phi_r(x_i) y_i}{\sum \phi_r^2(x_i)} ;$$

$r = 0, 1, \ldots k$

~~Generating orthogonal p~~

Generating Orthogonal Polynomials :

$$\phi_{r+1}(x) = 2(x - a_{r+1})\phi_r(x) - b_r \phi_{r-1}(x)$$

$\phi_0(x) = 1, \quad \phi_1(x) = 2(x - a_1)$

x is normalized so that $-1 \le x \le 1$

a_{r+1} & b_r are chosen as

$$a_{r+1} = \frac{\sum x_i \phi_r^2(x_i)}{\sum_{i=1}^{n} \phi_r^2(x_i)} ; \quad b_r = \frac{\sum \phi_r^2(x_i)}{\sum \phi_{r-1}^2(x_i)}$$

$r = 0, 1, \ldots k-1$

$b_0 = 0, \quad a_1 = \bar{x}$

$\hat{\gamma}_0 = \bar{y}$

H/W : Find the F-test for testing $\gamma_k = 0$.

Robust Regression :

- LSE's are the most efficient unbiased estimates of the regression coefficient when the errors are normally distributed.
- However, they are not very efficient when the dist^{n} of the errors in long-tailed.
  - ↳ we expect outlier in the data.
    [under normality assumption).
- When fitted a regression, we minimize, some average measure of the size of the residuals.

LS: Least mean of squares.

- fits a regression by minimizing the mean of the squared residuals.

$\Leftrightarrow$ the sum of the squared residuals.

$$\min_{\underline{b}} \frac{1}{n} \sum_{i=1}^{n} e_i^2(\underline{b}) \quad ; e_i(\underline{b}) = y_i - x_i' \underline{b}$$

Here average is interpreted as the mean and size as the square.

- The sensitivity of LS to outliers is due to two factors.

① If we measure size using the squared residuals, any residual with large magnitude will have a very large size relative to the others.

② By using a measure of location such as a mean that is not robust, any large square will have a very strong impact on the criterion.

Two remedies :

① Measure size in some other way

– by replacing e^2 by some f^n $\rho(e)$ which reflects the type of the residual in a less extreme way.

To be a sensible measure of size, the f^n ρ should be – symmetric $[\rho(e) = \rho(-e)]$

– non-ve $[\rho(e) \geq 0]$

– monotone $[\rho(|e_1|) \geq \rho(|e_2|)$ if $|e_1| \geq |e_2|]$

⟶ leads to M - estimations

② Replace the sum (or the mean) by a more robust measure of location such as the median or a trimmed mean.

↳ least median square (LMS)

→ least trimmed squares (LTS)

M-estimates:

Suppose the observed responces y_i independent and have the density f^ns.

$$f_i(y_i; \beta, \sigma) = \frac{1}{\sigma} f\left(\frac{y - x_i'\beta}{\sigma}\right) \quad \text{---(1)}$$

where σ is is a scale parameter.

Then the log likelihood corresponding to the density f^n is

$$l(\beta, \sigma) = -n\log\sigma + \sum_{i=1}^{n} \log\left[f\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right]$$

write $\rho = -\log f$

$$l(\beta, \sigma) = -\left\{n\ln\sigma + \sum_{i=1}^{n} \rho\left[\frac{y_i - x_i'\beta}{\sigma}\right]\right\}$$

So, to estimate β & σ using ML method, we minimize

$$n\log s + \sum_{i=1}^{n} \rho\left[\frac{e_i(b)}{s}\right].$$

Differentiating leads to the estimating eq.

$$\sum_{i=1}^{n} \psi\left(e_i(b)/s\right) x_i = 0 \quad \text{---(3)} \qquad \psi = \rho'$$

$$\sum_{i=1}^{n} \psi\left(e_i(b)/s\right) e(b) = ns. \quad \text{---(4)}$$

Example: $\rho(x) = \frac{1}{2}x^2$; $\psi(x) = x$.

(3) reduces to the normal equation of LSE,

$$\hat{\beta} = (x'x)^{-1}x'y$$

(4) $\rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2(\hat{\beta})$

(2) $\rho(x) = |x|$, the corresponding estimates are values of b & s. that minimise,

$$n \log s + \frac{1}{s} \sum_{i=1}^{n} |e_i(b)|$$

– is called the L_1 estimate.

- The L_1 estimate is the MLE if f is the double exponential density proportional to $\exp(-|x|)$
- An alt name is LAD (least absolute deviation) estimate.
- If we have no particular density in mind, we can choose ρ to the estimate by choosing a ρ for which $\psi = \rho'$ is bdd.
- We can generalize ③ & ④

$$\sum_{i=1}^{n} \psi\left[e_i(b)/s\right] \underline{x}_i = 0$$

$$\sum_{i=1}^{n} \chi\left[e_i(b)/s\right] = 0$$

where χ is also chosen to make the scale estimate robust.

The resulting estimates [soln of ③] is called the M-estimates. Since their definition is motivated by the ML estimating eq^{n}s.

- There is no requirement that ψ & χ be related to the density fn of the responses.

Example: Let, $\psi(x) = \begin{cases} -K & x < -K \\ x & -K \le x < K \\ K & x > K \end{cases}$

when K is a const to be chosen.

- The value of K usually chosen is 1.5, which gives a reasonable compromize between LS & L_1 estimate.
- It can be shown that a necessary cond^n for considering where when the parameter are estimated using (5) is

$$E[\psi(z)] = 0 \quad \text{---(6)}$$
$$E[\chi(z)] = 0 \quad \text{---(7)}$$

Considering
Weak: $\hat{\theta} \xrightarrow{P} \theta$
Strong: $\hat{\theta} \xrightarrow{a.s.} \theta$

when z has density fn f.

For example, (6) will be satisfied if f is symmetric about zero and of ψ is $\Rightarrow \psi(z) = \psi(-z)$.

Example: Suppose $\psi(x)$ is same as in Example (1). Then (6) is satisfied.

Then the scale parameter is (2) estimated by taking $\chi(x) = \psi^2(x) - c$ for some constant c.

Which is chosen to make the estimate consi. when f is the normal density.

From (7), we require that $c = E[\psi(z)^2]$, where z is standard normal.

Example: Another choice of χ is $\chi(z) = \text{sign}(|z| - 1/2)$ for some const c.

Then the corresponding estimating equation

$$\sum_{i=1}^{n} \text{sign}(|\psi_i(b)| - 1/c) = 0$$

$$\sum_{i=1}^{n} x\left(\frac{e_i(b)}{s}\right) = 0$$

- has soln.

$s = c \ \text{median}_i |e_i(b)|$

- mean absolute deviation (MAD estimate)

To make it consider, we require, $c^{-1} = \Phi^{-1}(3/4) = 0.6749$

$c = 1.4826$

- Regression coeff estimated using M-estimation method are almost as efficient as the LS method of the errors are normal but are much more robust if the error distn is long tailed.

- M estimates of regression coefficients are as vulnerable as LSE to outliers in the expl variables.

- Estimate Based on Robust location & scale measures

LMS: Least median Sq estimates.

Replacing the mean by a robust measure of location, but retain the squared residuals.

↳ LMS which minimizes

$\text{median}_i \, e_i^2(b)$.

An alternative to LMS estimator is to use the trimmed mean rather than the median.

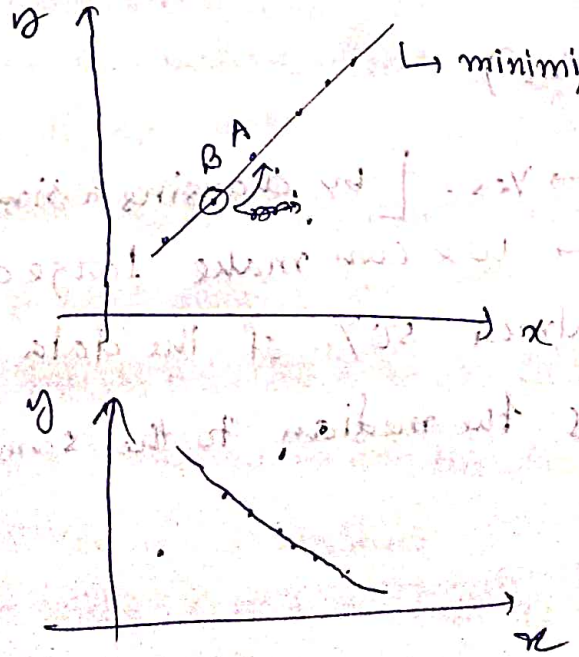
↳ LTS (least trimmed square) estimates.

This minimizes $\sum_{i=1}^{h} e_i^2(\underline{b})$

where h is chosen to achieve a robust estimator and $e_{(1)}^2(\underline{b}) \leq e_{(2)}^2(\underline{b}) \leq \ldots \leq e_{(n)}^2(\underline{b})$ are ordered squared residuals.

The amount of trimming has to be quite severe to make robust.

- $h = [n/2]+1$ is a popular choice → which amounts to trimming of 50% of residuals.
- LMS & LTS are very robust to outliers in both errors and exp variables but can be very unstable in a different way.



- In addition, these estimates are very inefficient compared to LSE of the data are actually normally distributed.
- Asymptotic relative efficiency of LMS. relative to LSE is zero. $\left[\frac{V(LSE)}{V(LMS)} \to 0 \text{ as sample size} \to \infty \right]$.

and $\frac{Var(LSE)}{Var(LTS)} \approx 0.08$.

Measuring Robustness:

Breakdown Point (BD): measure how well an estimate can resist gross corruption of a fraction of the data.

Suppose that we select a fraction of the data. Can we cause an arbitrarily large change in the estimate by making a suitable large change in the selected data pts.

(Sometimes Yes)

Sample mean → Yes. [by changing a single data pt

" median → we can make large changes to almost 50% of the data without changing the median to the same extent.

Def^n: The finite sample breakdown pt of an estimate is the fraction of the data that can be given arbitrary values without making the estimator arbitrarily .

- The sample mean has BD pt $1/n$ and the sample median a BD pt of almost $1/2$.
- A BD point of $1/2$ is the best possible. If more than 50% of the sample is contaminated, it is impossible to distinguish between "good" or "bad" obs. Since the outliers we now typical obs of the sample.

- LSE of a regression coefficient is a linear comb^n of the responses

BD pt of LSE is $1/n$

$\sum |y_i - \theta|$ w.r.t. θ Least absolute derivation estimator $\sum_{i=1}^{n} |y_i - x_i'\beta|$

- Since the median has a very high BD point, and median of the data $y_1, \dots, y_n$ minimizes the Least absolute deeviation $\sum |y_i - \theta|$ as a fn coefficients would also have a high BD point.
- This is not the case, in fact, the BD pt of L_1 estimator is the same as the LSE.
- It can be show that
  - when x is a of full rank there is a value of $\underline{b}$ minimizing $\sum_{i=1}^{n} |e_i(\underline{b})|$ for which at least p

residuals are zero.

- Further of one data point is arbitrarily far from the other this data pt must have a zero residual.

- It follows that by moving the data pt can arbitrary amount, we must also be moving the fitted planee by an arbitrary amount, since the fitted plane passes through the extreme data pt.

- The same is true for M-estimator.

- The LMS & LTS estimators were inefficient compared to M-estimator.

But they have BD points almost 1/2, the best possible one.

Influence Curve

F: k dim distⁿ fⁿ.

θ: poplⁿ parameter that depends on F.

we write $\theta = T(F)$

we call T as a statistical functional.

Example: Simplest example of statistical functional is the mean $E_F(x)$ of a r.v. X.

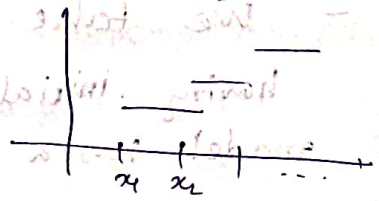
$$T(F) = E_F(x) = \int x \, dF(x)$$

Example: $\underline{z}$ is a random vector with distⁿ fⁿ f, then $E(\underline{z}\underline{z}') = \int \underline{z}\underline{z}' \, dF(\underline{z})$

Def^n: If $x_1, x_2, \ldots, x_n$ are iid RVs each with dist^n f^{n} f, then the empirical dist^n f^{n} (edf) F_n ($\hat{F}_n$) is the dist^n f^{n} which places n^{-1} at each of the pts x_i, $i = 1, 2, \ldots, n$.

$$F_n(t) = \frac{\text{no of elements in the sample} \leq t}{n}.$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(x_i \leq t)$$



Integration w.r.t. edf is just the averaging. If h is a function then

$$\int h(x)\, dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} h(x_i)$$

– Many statistic used to estimate parameters $T(F)$ are the plug-in-estimator of the $T(F_n)$, where F_n is the edf based on a sample of size n.

Example: $\underline{Z}_1, \underline{Z}_2, \ldots, \underline{Z}_n$ random sample from a multivariate dist^n F. The plug-in-estimator of

$$T(F) = \int \underline{z}\, dF(\underline{z})$$

as $T(F_n) = \int \underline{z}\, dF_n(\underline{z}) = n^{-1} \sum_{1}^{n} \underline{z}_i \rightarrow$ the sample mean.

11^{ly}, the plug-in estimator of

$$T(F) = \int \underline{z}\underline{z}'\, dF(\underline{z}) \text{ is}$$

$$T(F_n) = \int \underline{z}\underline{z}'\, dF_n(\underline{z}) = \frac{1}{n} \sum_{i=1}^{n} \underline{z}_i \underline{z}_i'$$

- Conisder a regression with a response variable Y and exp variables $x_1, \ldots, x_{p-1}$.

- Regard the regression data, $(\underline{x}_i, y_i)$; $i = 1, 2, \ldots, n$ as n iid random $(p+1)$ vectors, distributed as $(\underline{x}, y)$, having a joint dist^n F.

- We take the vector $\underline{x}$ is random and having initial element as 1; if the regression model has a constant term.

- We also assume that the conditional dist^n $y|\underline{x}$ has density f^n $g\left(\frac{y - \underline{x}'\underline{\beta}}{\sigma}\right)$; where g is a known density.

(LS): Consider the functional

$$T(F) = \left[E_F(\underline{x}\underline{x}')\right]^{-1} E_F[\underline{x}\, y]$$

The plug-in-estimator of this T is

$$T(F_n) = \left[n^{-1}\sum \underline{x}_i \underline{x}_i'\right]^{-1} \left[n^{-1}\sum \underline{x}_i y_i\right]$$

$$= (X'X)^{-1} X'\underline{y}$$

- Suppose that F is a dist^n f^n.

- We can model a small change in F at a fixed (non-random) value $\underline{z} = (\underline{x}_0', y_0)$. by considering a mixture of dist^n.

$$F_t = (1-t)F + t\delta_{z_0}$$

Where δ_{z_0} is the dist^n f^n of the constant z_0 and $t \geq 0$ is close to zero.

The sensivity of T can be measured by the rate at which $T(F_t)$ changes for small values of t.

Def^n: The influence curve (IC):

of a statistical functional T is the derivative of $T(F_t)$ w.r.t. t evaluated at $t=0$. and is a measure of the rate at which T responds to a small change amount of contamination at z_0.

IC depends on both F & z_0

$$IC(F, z_0) = \left.\frac{dT(F_t)}{dt}\right|_{t=0}$$

Example: Let, $T(F) = \int x\, dF(x)$.

$$T(F_t) = \int x\, dF_t(x)$$
$$= (1-t)\int x\, dF(x) + t\int x\, d\delta_{x_0}(x)$$
$$= (1-t)\cdot T(F) + t x_0$$

So, $$\frac{T(F_t) - T(F)}{t} = x_0 - T(F)$$

$$IC(F, x_0) = x_0 - T(x)$$

↳ un bdd in x_0

This suggests that a small amount of contamination can cause an arbitraily large change.

→ The term is highly non-robust.

Example: (IC for LSE)

Let, T be the LSE functional

$$\left[E_F\left[\underline{x}\,\underline{x}'\right)\right]^{-1} E_F(\underline{x}\,\underline{y})$$

Write $\Sigma_F = E_F(\underline{x}\,\underline{x}')$ & $\gamma_F = E_F(\underline{x}, \underline{y})$

Then $T(F_t) = \left(\Sigma_{F_t}\right)^{-1} \gamma_{F_t}$

$$\Sigma_{F_t} = E_{F_t}(\underline{x}\,\underline{x}') = (1-t) E_F(\underline{x}\,\underline{x}') + t\,\underline{x}_0\,\underline{x}_0'$$

$$= (1-t)\Sigma_F + t\,\underline{x}_0\,\underline{x}_0'$$

$$= (1-t)\left\{\Sigma_F + t'\underline{x}_0\,\underline{x}_0'\right\}$$

$$t' = \frac{t}{1-t}$$

$$\Sigma_{F_t}^{-1} = (1-t)^{-1}\left[\Sigma_F^{-1} - \frac{t'\Sigma_F^{-1}\underline{x}_0\underline{x}_0'\Sigma_F^{-1}}{1+t'\underline{x}_0'\Sigma_F^{-1}\underline{x}_0}\right]$$

$$= (1-t)^{-1}\left[\Sigma_F^{-1} - \frac{t'\Sigma_F^{-1}\underline{x}_0\underline{x}_0'\Sigma_F^{-1}}{1+o(t)}\right]$$

$$\gamma_{F_t} = (1-t)\gamma_F + t\underline{x}_0 y_0$$

$$T(F_t) = T(F) + t'\Sigma_F^{-1}\underline{x}_0 y_0 - t'\Sigma_F^{-1}\underline{x}_0\underline{x}_0'\Sigma_F^{-1}\gamma_F + o(t)$$

$$\left[x_t = o(t) \Rightarrow \frac{1}{t}x_t \to 0 \text{ as } t \to 0\right]$$

$$\Rightarrow \frac{T(F_t) - T(F)}{t} = \Sigma_F^{-1}\underline{x}_0 y_0 - \Sigma_F^{-1}\underline{x}_0\underline{x}_0'\Sigma_F^{-1}\gamma_F + o(1)$$

Take $t \to 0$;

$$IC(F, \underline{x}_0) = \Sigma_F^{-1}\underline{x}_0\left[y_0 - \underline{x}_0' T(F)\right]$$

→ is unbounded in both x_0 & y_0 which indicate that LSE is not robust.

Stepwise Regression:

Suppose that we have the model

$$\underline{Y} = X\underline{\beta} + \underline{\epsilon}$$

X is $n \times (k+1)$ want to identify the 'significant' variables having non-zero regression coefficients.

Suppose, we divide the k variable into two groups.

1. The 1st one consists of $p-1$ variables that we consider as important.
2. The second are which contains $k-p+1$ variables consists of variables whose coefficient we ~~sus~~ suspect are zero.

We can test if the second set contains no significant variable using

$$F = \frac{RSS_p - RSS_{k+1}}{RSS_{k+1}} \cdot \frac{n-k-1}{k-p+1}$$

- test which discriminate between two models having $p-1$ and k variables.

If $k=p$; then

$$f^* = \frac{RSS_p - RSS_{p+1}}{RSS_{p+1}} (n-p-1) \sim F_{1, n-p-1}$$

— If addition of a specified extra variable is necessary.

Forward Selection:

1. Stand with the model with constant term by only compute F^* with $p=1$ for all the available exp variables and pick that variable for which F^* is the largest.
2. Repeat the procedure for $p=2,3,\ldots$ selecting at each step the variable currently not included that gives the maximum value of F^*.
3. Step of the maximum value of p^* at any stage does not exceed some threshold value f_{IN}.
4. This procedure is called forward selection.

A slight variation on FS is to pick up at each stage the variable not currently included in the model having the largest ~~possible~~ partial correlation with the response given the variable currently included.

Backward elimination:

As an alternative to FS, we can start the full model using all K variables (provided $K+1<n$) and compute F with $p=K$ for each of the K variables.

- Eleminate the variable having the smallest F-statistic from the model provided F is less than some thresold value.
- The procedure is continued untill all variable are eliminate or the smallest F fails to be less than F_{out}.

- This procedure is called the backward elimination.

Stepwise Regression

A method that combines FS and BE.

→ This is just a FS followed by a BE step at each stage.

- Start with a constant term alone then perform FS step, adding a single variable.
- This is followed by a BE step, removing a variable of the corresponding f is less than f_{out}
- This combination of an FS step followed by a BS step is replaced until no further variable is added at the FS step. Provided $f_{out} \leq f_{IN}$ the stepwise regression algorithm must eventually terminates.
- Suppose that at the beginning of an FS step there are p-1 variables in the current model, which has a residual RSS_p as RSS_{p-1}.

Then either, no variable can be added, and the algorithm terminates or an additional variable is added, resulting in the new RSS_{p+1} which must satisfy.

$$\frac{RSS_p - RSS_{p+1}}{RSS_{p+1}} (n-p-1) > f_{IN}$$

or, $RSS_p > RSS_{p+1}\left(1+\frac{f_{IN}}{n-p-1}\right) > RSS_{p+1}$

Now, do a BE step. Either no variable is deleted or we get a new model with $p-1$ variables and a new RSS equal to RSS_{ap}^*

$$RSS_p^* < RSS_{p+1}\left(1+\frac{F_{out}}{n-p-1}\right) < RSS_{p+1}\left(1+\frac{f_{IN}}{n-p-1}\right)$$

$$\leq RSS_p$$

If $F_{out} \leq f_{IN}$.

Testing for Heteroscedasticity : Breusch-Pasan Test

The linear regression equation with heteroscedastic error

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} + \epsilon_i \; ; \; i = 1, 2, \ldots n.$$

$E(\epsilon_i) = 0$

$V(\epsilon_i) = \sigma_i^2$

Assume that error variables is of the linear term of independent variables,

$$\sigma_i^2 = b_0 + b_1 x_{i1} + \cdots + \delta_{k-1} x_{ik-1} \; ; \; k : \text{some exp variables}$$

Choose the set of $k-1$ variables with which we can assume that σ_i^2 vary..

H_0: $Var(\epsilon_i | \underline{x}_i) = \sigma_i^2$ $\quad H_1$: not H_0.

$\Leftrightarrow$ H_0: $\delta_1 = \cdots = \delta_{k-1} = 0$.

H_1: not all δ's are zero

• Since we never know the actual error, we use then estimates $\hat{\epsilon}_i$, which are OLS residuals.

consider $\epsilon_i^2 = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_{k-1} x_{i,k-1} + v_i$

We can use a F-test.

- f statistic for testing this hypothesis is the goodness of fit test.

Let, $R^2_{\hat{\epsilon}^2}$ be the multiple ~~correction~~ correlation coeff of this regression.

$$F = \frac{R^2_{\hat{\epsilon}^2}}{1 - R^2_{\hat{\epsilon}^2}} \cdot \frac{n-k}{n-1} \sim f_{n-1, n-k} \quad \text{under } H_0.$$

Reject H_0 if $F > f_{n-1, n-k}(\alpha)$

A form of ~~Breusch~~ Breusch-Pagan test for heteroscedasticity is $n \times R^2_{\hat{\epsilon}^2} \sim \chi^2_{k-1}$ [$k-1 \rightarrow$ no of indpt variable]

(Lagrange's Multiplier Statistic)

$$LM = \left(\frac{\partial l}{\partial \theta}\right)' \left[-E\left[\frac{\partial^2 l}{\partial \theta \partial \theta'}\right]\right] \left(\frac{\partial l}{\partial \theta}\right)$$

BP test has been shown to be sensitive to any violation of normality assumption.

Auxiliary equation for error variance can be

$$\sigma_i^2 = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_{k-1} x_{ik-1}$$

$$\sigma_i = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_{k-1} x_{ik-1}$$

$$\ln(\sigma_i^2) = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_{k-1} x_{ik-1}$$

or $\sigma_i^2 = \exp[\delta_0 + \delta_1 x_{i1} + \cdots + \delta_{k-1} x_{ik-1}]$

We donot know σ_i^2, estimate $\hat{\epsilon}_i^2$ and use $\hat{\epsilon}_i^2$ for σ_i^2. $|\hat{\epsilon}_i|$ for σ_i and $\ln(\hat{\epsilon}_i^2)$ for $\ln(\sigma_i^2)$

White test; Holdfeldt - Quandt Test; Breusch Godfrey test

White test: This is explicitly intended to test for forms of heteroscedasticity.

- The relation of ϵ^2 with all the independent variables (x_c), the squares (x_i^2) and all cross-product ($x_i x_j$, $i \neq j$)
- Just like BP test, we consider $\hat{\epsilon}_i^2$ on all this above variables and compute $R^2_{\hat{\epsilon}^2}$. Let the total no of $(x_i, x_j^2, x_i x_j)$ be k.
- $n R^2_{\hat{\epsilon}^2}$ is distributed as χ^2_k for large sample.

The abundance of independent variables is a weakness in this original of white test.

The idea is to derive by observing the original regression model.

- Consider in the 2nd stage.

$$\hat{\epsilon}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + v$$

where $\hat{y}$ is the producted value of y.

to

Use f or LM test $H_0: \delta_1 = \delta_2 = 0$

[very bad for small sample]

→ a sp. case of white test

→ perform very badly for small samples.

The holdfeldt Quandt Test:

- White's test and BP test both focus on smoothly changing variances for the error term.

- The GQ test for heteroscedasticity divide n observations into graphs and tests whether two specific groups have disturbances with the same variances, because the errors are unobserv we compare the MSEs for this groups.

Test:

1. Divide n observations into h groups. Let, $n_1, n_2, \ldots, n_h$ denote the no of observation in groups $1, 2, \ldots h$

2. Choose two groups 1, 2 say for which to test $H_0: \sigma_1^2 = \sigma_2^2$ v/s $H_1: \sigma_1^2 = \sigma_2^2$

3. Estimate K, explanator model of interest by OLS using n_1 obs in group 1 Let RSS be SSR,

4. Do it for group - 2 , let it be SSR_2

5. Name the larger one of $\frac{SSR_1}{n_1-k}$ and $\frac{SSR_2}{n_2-k}$ as $\frac{SSR_2}{n_L-k}$ and the smaller one as $\frac{SSR_S}{n_S-k}$

6. Form the statistic

$$G = \frac{SSR_L/(n_L'-k)}{SSR_S/(n_S-k)}$$

7. Reject H_0 if $G > F_{n_L-k, n_S-k}^{\alpha}$

8. The GQ test is valid if the underlying errors are normally distributed.

9. It is approximately valid if both n_L-k and n_S-k are large and errors are not normally distributed.

. The most frequent approaches of GQ test involves dividing the sample in two groups.

Consider $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ $\quad V(\epsilon_i) = \sigma_i^2$.

$Cov(\epsilon_i, \epsilon_j) = 0$, $i \neq j$

how can you estimate $var(\hat{\beta}_1|x)$

$$var(\hat{\beta}_1|x) = \frac{\sum (x_i-\bar{x})^2 \sigma_i^2}{\left[\sum (x_i-\bar{x})^2\right]^2}$$

$\hat{\beta}_1$ ↓ OLS

$$\neq \frac{\sigma^2}{\sum (x_i-\bar{x})^2}$$

$$\hat{\beta}_1 = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}$$

$$= \frac{\sum (x_i-\bar{x}) y_i}{\sum (x_i-\bar{x})^2}$$

Breusch-Godfrey Test for AR(q):

$$y_t = \beta_0 + \beta_1 x_t + e_t$$

$$e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \cdots + \rho_q e_{t-q} + \epsilon_t$$

$\epsilon_t \sim$ iid $N(0, \sigma^2)$

or uncorrelated.

H_0: $\rho_1 = \rho_2 = \cdots = \rho_q = 0$

Estimation β_0 and β_1 using OLS and obtain $\hat{e}_t$ (residuals)

Regress the estimated residuals on the lagged values upto ~~order~~ lag q and all original variables.

$$\hat{e}_t = \rho_1 \hat{e}_{t-1} + \cdots + \rho_q \hat{e}_{t-q} + \gamma_1 x_t + \gamma_0 + \epsilon_t \quad t = q+1, \ldots$$

↳ Treat it as an regression model and obtain R^2

$NR^2 \sim \chi^2_q$ [asymptotic approximation]

N: no of data pts available for the second stage regression

$(n-q)\, R^2_{\hat{e}} \sim \chi^2_q$

Ridge Regression :

In general case, we need a reliable method of estimating the optimal value of k if ridge estimate is to be superior to LSE.

- Popular method is based on CV or GCV.

The estimate of the prediction error of the ridge predictor

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i - x_i'\hat{\beta}_{(-i)}(k)\right]^2 .$$

$\hat{\beta}_{(-i)}(k)$: ridge estimate to β calculated ridge coeff k but leaving out the ith data point.

Simplify it $\frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - x_i'\hat{\beta}(k)}{1-a_{ii}(k)}\right]^2$

Where, $a_{ii}(k)$: ith diagonal element of $A(k)$

$$= X(X'X + kI_p)^{-1}X'$$

The value of k is then chosen to minimise this criterion.

- The ridge estimate can be regarded as the soln of a constrained LS problem.

Consider minimising the SS

$$\|y - Xb\|^2$$

subject to the constraint $\sum b_j^2 \le s$, where s is some specified constant.

- If $\sum \hat{\beta}_j^2 \le s$, the LSE solves this constrained prblm

If $\sum \hat{\beta}_j^2 > s$, then the soln is different and given by the Lagrange's Multiplier.

$$X'X\underline{b} - X'y + \lambda \underline{b} = \underline{0} \quad ; \quad \sum b_j^2 = s.$$

which has the soln of the form

$$\hat{\underline{\beta}}(\lambda) = (X'X + \lambda I_p)^{-1} X'y$$

where λ satisfies $\sum \hat{\beta}_j^2(\lambda) = s$.

- If $X'X = I_p$, the ridge estimate is

$$\hat{\underline{\beta}}(\lambda) = \frac{\hat{\beta}}{1+\lambda} \rightarrow$$ therefore, each coefficient is shrunk by a constant factor.

Garrote & Lasso Estimate :

- One drawback of the ridge estimate is unlike the subset selection it retains all the variation in the model. No possibility of a simpler model with fewer variables.

An alternative to ridge to set a simple model Garrote.

In Garrote, the individual LS coefficients $\hat{\beta}_j$ are shrunk by non-ve quantity c_j

- leads to garrote estimate

$\tilde{\beta}_j = c_j \hat{\beta}_j \rightarrow$ LSE

↑

shrinkage factors.

The c_j's are chosen to minimise the LS estimators

$$\|\underline{Y} - X\underline{\tilde{\beta}}\|^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p-1} x_{ij}\, c_j\, \hat{\beta}_j \right)^2$$

subject to the constraints $c_j \geq 0$, $j = 0, 1, 2, \ldots, p-1$

& $\sum_{j=0}^{p-1} c_j \leq s$

Where s is some specified positive constant.

Date: 23, 12, 2021

for $s > p$, the choice $c_j = 1$, $j = 0(1)p-1$ gives $\tilde{\beta}_j = \hat{\beta}_j$, unconstrained minimum

As s is reduced, the shrinkage coefficients get smaller and some are even forced to zero

Thus, garrote is regarded as a compromise between the ridge and the subset selection

- The problem of finding the shrinkage coefficients c_j can be reduced to a standard LS problem (constrained)

- write $\|Y - X\tilde{\beta}\|^2 = \|Y - X\hat{\beta} + X\hat{\beta} - X\tilde{\beta}\|^2$

$= \|Y - X\hat{\beta}\|^2 + \|X\hat{\beta} - X\tilde{\beta}\|^2 + 2(Y - X\hat{\beta})'(X\hat{\beta} - X\tilde{\beta})$

$= RSS + \|X\hat{\beta} - X\tilde{\beta}\|^2 + 0$

Let $a_{ij} = x_{ij}\hat{\beta}_j$, $A = ((a_{ij}))$, $\underset{\sim}{c} = (c_0, \cdots, c_{p-1})'$

$X\tilde{\beta} = A\underset{\sim}{c}$, $\underset{\sim}{d} = X\hat{\beta}$

$\Rightarrow \|Y - X\tilde{\beta}\|^2 = RSS + \|\underset{\sim}{d} - A\underset{\sim}{c}\|^2$ ——— (*)$_2$

* the vector $\underset{\sim}{c}$ which minimizes (*), subject to the constrain $c_j \geq 0$, $j = 0(1)p-1$ & $\sum c_j \leq s$ is the same as the vector $\underset{\sim}{c}$ that minimizes (*)$_2$, so

$\|A\underset{\sim}{c} - \underset{\sim}{d}\|^2$ subject to the same constraints

- A version of garrote which drops the requirement that the shrinkage coefficients be non-ve & satisfy $\sum \beta_j^2 \leq s$ similar to the ridge

Breiman : non-garrote $\longrightarrow$ $\sum c_j \leq s$

garrote $\longrightarrow$ $\sum \beta_j^2 \leq s$

LASSO

Date : / /

consider a different estimator which minimizes

$$\sum_i (y_i - x_i'\beta)^2 + \lambda \sum_{j=0}^{p} |\beta_j| \quad , \lambda \rightarrow \text{tuning parameter}$$

The only difference from RR being that absolute values instead of squares are used in the penalty ()

- The change in penalty is small, but the impact on the estimator is huge
- Like RR, penalizing the absolute values of the coefficients introduces shrinkage.
- However, unlike RR, some of the coefficients are shrunken all the way to zero
  - such solutions with multiple values that are identically equal to zero, are said to be sparse
- The penalty thereby performs a sort of variable selection

The resulting estimator was thus named LASSO for "Least Absolute Shrinkage and Selector Operator"

Note: centered and scaled model: $Y = X_S \beta + \varepsilon$, $X_S = (1, X^*)$

if β has a $N_p(\mu, \sigma^2 V)$ prior distribution $\beta = (\beta_0, \beta^*)'$

then the posterior mean of β is $(X_S'X_S + V^{-1})(V^{-1}\mu + X_S'Y)$

if $\mu = 0$ and $V = \begin{pmatrix} c^{-1} & 0 \\ 0 & k^{-1} I_{p1} \end{pmatrix}$

the posterior mean (the Bayes estimate) of Y is exactly the ridge estimate

Generalized Linear Model (GLM)

- GLMs extended ordinary regression models to encompass non-normal response distribution and modelling functions of the mean
- They have three components:
  1. A random component, identifies the response variable Y and its probability distribution
  2. A systematic component, specifies explanatory variables used in a linear predictor ()
  3. A link (), specifies the function $E(Y)$ that the model

Date : / /

- The random component of a GLM consists of a response variable $\underset{\sim}{Y}$ with independent observations $(Y_1, Y_2, \ldots, Y_n)$ from a distribution in the natural exponential family :

$$f(y_i; \theta_i) = a(\theta_i) b(y_i) \exp\{\theta_i Q(\theta_i)\}$$

- the value of the parameter θ_i varies for $i = 1(1)n$ as a ( ) of the values of the explanatory variables.
- the parameter $Q(\theta)$ is called the natural parameter
- the systematic component of a GLM relates a vector $(\eta_1, \eta_2, \ldots, \eta_n)'$ to the explanatory variable through a linear model
- let x_{ij} denote the value of explanatory variable j $(j = 0(1)p-1)$ for subject i

$\eta_i = \sum_{j=0}^{p-1} x_{ij} \beta_j$; $i = 1(1)n$ is called the linear predictor

- usually $x_{0i} = 1 \; \forall i$, representing the coefficient of an intercept term β_0 in the model
- the third component of a GLM is a link function that connects the random component and the systematic component
- let $\mu_i = E(Y_i)$; $i = 1(1)n$
- the model links μ_i to η_i by $\eta_i = g(\mu_i)$ where the link function g is a monotonic, differentiable function

$$g(\mu_i) = \sum_{j=0}^{p-1} x_{ij} \beta_j \quad ; \; i = 1(1)n$$

- $g(\mu) = \mu$ ⇒ identity function $\eta_i = \mu_i$
  - → specifies linear model for the mean itself
  - → link for ordinary regression with normally distributed Y
- the link function that transforms the mean to the natural parameter is called the canonical link

Date : / /

- We present the "success" and "failure" outcomes by 1 and 0
- A bernoulli trial has possibilities : $P(Y=1)=\pi$, $P(Y=0)=1-\pi$
  So, $E(Y)=\pi$

  The probability mass function $f(y;\pi)=\pi^y(1-\pi)^{1-y}$
  $=(1-\pi)\left(\frac{\pi}{1-\pi}\right)^y$

  $\Rightarrow f(y;\pi)=(1-\pi)\exp\{y\log\left(\frac{\pi}{1-\pi}\right)\}$, for $y=0,1$

  $\theta=\pi$, $a(\pi)=1-\pi$; $Q(\pi)=\log\left(\frac{\pi}{1-\pi}\right)$; $b(y)=1$

The natural parameter $\log\left(\frac{\pi}{1-\pi}\right)$ is the log odds of response outcome 1, the logit of π

This is the canonical link function
↳ GLM with logit link are referred to as logistic regression models or logit models

Poisson loglinear model for count data

$$f(y;\mu)=\frac{e^{-\mu}\mu^y}{y!}=\frac{1}{y!}\exp\{-\mu+y\log(\mu)\}$$

$\therefore \theta=\mu$; $a(\theta)=e^{-\mu}$; $b(y)=\frac{1}{y!}$, $Q(\theta)=\log(\mu)$

So the natural parameter is $\log\mu$ and canonical link function is the log link , $\eta=\log(\mu)$

$\log(\mu_i)=\sum_{j=0}^{p-1}x_{ij}\beta_j$; $i=1,\dots,n$

↳ Poisson log linear model

Deviance of a GLM / Deviance function :

- for a particular GLM with observations $(y_1,\dots,y_n)$, let $l(\hat{\mu};y)$ denote the log-likelihood ( ) expressed in terms of the mean
- let $l(\hat{\mu};y)$ denote the maximum of the log-likelihood for the model
- consider for all possible models, the maximum achievable log-likelihood is $L(y;y)$
  ↳ this occurs for the most general model having a separate parameter for each observation and

↳ called the "saturated model"
(contains as many parameters as there are data points)

Date :

- this is not a useful model as it doesn't provide any data reduction
- however, it serves as a baseline for comparison with other model fits

- the deviance ( ) is defined/based on likelihood ratio as:

$$D = -2 \log_e \left[\frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}} \right]$$

$$\Rightarrow D = -2 \log_e \left[l(\hat{\mu}, y) - l(y, y) \right]$$

- this is the LR statistic for testing the null hypothesis that the model under consideration holds against the general alternative (i.e., the saturated model)

deviance is used for model checking and for inferential comparison of models.

GLM for Binary Data

y: binary response variable (success, failure)

Each obs has one of the two outcomes, denoted by 1 and 0. → use as a trial for a single Bernoulli trial

$E(y) = P(Y=1) = \pi(\underline{x})$, say, reflecting its value of $\underline{x} = (x_1, \ldots, x_p)$ of exp variables.

$Var(y) = \pi(\underline{x})(1-\pi(\underline{x}))$

Linear Probability Model

$\pi(\underline{x}) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p$.

$\pi(\underline{x}) = E(y|\underline{x})$

↳ a linear prob model

↳ with independent obs, it is a GLM with Binomial RC and identity link.

- The model has a structural problem

$\pi(\underline{x}) \in [0,1]$, whereas the linear predictor takes values in the entire real line.

$\pi(\underline{x}) < 0$ and/or $\pi(\underline{x}) > 1$ for some values of $\underline{x}$.

So the model is valid on a very restricted region of $\underline{x}$ values.

- β_j is the change in $\pi(\underline{x})$ for one unit change in x_j

↳ constant

so the OLSE's will not be optimal. So ML estimator is more efficient than OLSE's.

→ Madly unsatisfactory $\hat{\pi}(x)$ as an estimator.

Also, y, being binary, is very far from normality, so the usual t-test, F-test sampling distn for LSE's do not apply

Logistic Regression Model

- Usually binary data result from a unknown relationship between $\pi(x)$ and x.
  - A fixed change in x often has less impact when $\pi(x)$ is near 0 or 1, than when $\pi(x)$ is near 0.5.
  - In practice, non-linear relationship between $\pi(x)$ and x are often monotone with $\pi(x)$ increasing continuously or decreasing continuously as x decreases
    - ↳ The S-shaped curve are typically observed

The important curve with this shape is the model

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

↳ Logistic regression model with one exp variable.

As $x \uparrow$, $\pi(x) \uparrow$ when $\beta > 0$ and $\downarrow$ when $\beta < 0$.

General Form : $Y_i | x_i \sim$ Bernoulli (π_i) , $\pi_i = \pi(x_i)$
$= P(Y_i = 1 | x_i)$

$$\text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_h x_{ih}$$

$$= \beta_0 + \underline{x}_i' \underline{\beta}$$

$$\underline{\beta} = (\beta_1, \ldots, \beta_h)'$$

$\beta_0 \rightarrow$ intercept $\underline{\beta} \rightarrow$ vector of parameters corresponding to the independent variables.

$$\eta_i = \beta_0 + \underline{\beta}' \underline{x}_i$$

- In general all odds are defined as the ratio of probability of $Y=1$ to that of $Y=0$.

Note: To compute $\beta' \underline{x}_i$, it requires $\underline{x}_i$ to be a vector of numeric values.

- It does not directly apply to categorical variables.

- For a binary covariate, we can represent the differential effect assisted by the two levels using a binary indicator or dummy variable, taking values 0 and 1.

- For categorical variable with k levels ($k > 2$), we may designate one level as a reference and use $k-1$ binary indicators to represent the individual difference for each of the remaining $k-1$ levels

the name of the logistic regression model.

$$F(x) = \frac{\exp(x)}{1+\exp(x)} \quad -\infty < x < \infty$$

↳ cdf of the standard logistic r.v.

• If Y_i is the no of subjects with the event and n_i is the size of the stratum for the ith unique value of x_i, then $Y_i \sim \text{Binomial}(n_i, \pi_i)$, π_i is the prob of success.

Binomial GLM for 2×2 contingency table

• For a binary response, we have a single exp variable that is also categorical. Values are 0 and 1.

y \ x	0	1
0	n_{00}	n_{10}
1	n_{01}	n_{11}

For a given link fn.

$$\text{link}[\pi(x)] = \alpha + \beta x.$$

as the effect of x is described as

$$\beta = \text{link}[\pi(1)] - \text{link}[\pi(0)]$$

• For the identity link, $\beta_x = \pi(1) - \pi(0)$ is the difference between proportions.

• For log link, $\beta = \log\left(\frac{\pi(1)}{\pi(0)}\right)$ is the log of relative risk.

• For logit link, $\beta = \text{logit}(\pi(1)) - \text{logit}(\pi(0))$

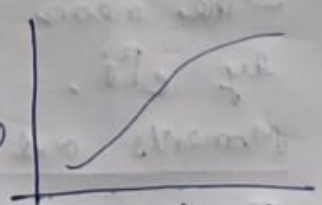
$$= \log \frac{\pi(1)}{1-\pi(1)} - \log \frac{\pi(0)}{1-\pi(0)}$$

$$= \log \left[\frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))} \right]$$ → is the log odds ratio.

Probit and inverse cdf link function

• This suggests a model for binary response having form

$\pi(x) = F(x)$ for some cdf



$$\pi(x) = \Phi(\alpha + \beta x)$$

- When Φ is strictly increasing over the entire real line its inverse exists and

$$\Phi^{-1}(\pi(x)) = \alpha + \beta x.$$

So the link fn for the GLM is Φ^{-1}.

- This curve has the shape of normal cdf when Φ is $N(0,1)$ → called Probit model.
- When $\beta > 0$, the logistic regression curve is a cdf for the logistic distn.
- When $\beta < 0$, the curve for $1 - \pi(x)$ has that appearance

The cdf of the logistic distn with mean μ and dispersion parameter $\tau > 0$, then

$$F(x) = \frac{\exp[(x-\mu)/\tau]}{1 + \exp[(x-\mu)/\tau]}, \quad -\infty < x < \infty.$$

Poisson Log-linear model

$$\log \mu(x) = (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

$$\Rightarrow \mu(x) = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

$$= e^{\beta_0}(e^{\beta_1})^{x_1}(e^{\beta_2})^{x_2} \cdots (e^{\beta_k})^{x_k}$$

- A 1-unit increase in x_j has a multiplicative impact of e^{β_j}
- The mean at $x_j + 1$ equals the mean at x_j multiplied by e^{β_j}.

Moments and likelihood for GLM

The exponential dispersion family

$Y_1, \cdots, Y_n$ are independent with pdf or pmf for Y_i

$$\frac{\partial L_i}{\partial \theta_i} = L_i = [y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i;\phi)$$

$$\frac{\partial L_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)}, \quad \frac{\partial^2 L_i}{\partial \theta_i^2} = -\frac{b''(\theta_i)}{a(\phi)}.$$

$b'(\theta)$, $b''(\theta)$ > 1st and 2nd ~~di~~ derivative of $b(\theta)$.

- Applying general likelihood results

$$E\left[\frac{\partial L}{\partial \theta_i}\right] = 0, \quad -E\left[\frac{\partial^2 L}{\partial \theta^2}\right] = E\left[\left(\frac{\partial L}{\partial \theta}\right)^2\right]$$

This holds under regularity conditions satisfied by the exponential family

$$E\left[\frac{y_i - b'(\theta_i)}{a(\phi)}\right] = 0 \Rightarrow \mu_i = E(y_i) = b'(\theta_i)$$

$$\frac{b''(\theta_i)}{a(\phi)} = E\left[\left\{\frac{y_i - b'(\theta_i)}{a(\phi)}\right\}^2\right] = \frac{\text{var}(y_i)}{\{a(\phi)\}^2}$$

$$\Rightarrow \text{var}(y_i) = b''(\theta_i)\, a(\phi)$$

The fn $b(\cdot)$ determines the moments of y_i

- When y_i is Poisson

$$f(y_i;\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \exp[y_i \log \mu_i - \mu_i - \log y_i!]$$

$$= \exp[y_i\theta_i - \exp(\theta_i) - \log y_i!]$$

When $\theta_i = \log \mu_i$, $b(\theta_i) = \exp(\theta_i)$, $a(\phi) = 1$,

$$\therefore c(y_i;\phi) = -\log y_i!$$

$$E(y_i) = b'(\theta_i) = \exp(\theta_i) = \mu_i$$

$$\text{var}(y_i) = b''(\theta_i) = \exp(\theta_i) = \mu_i.$$

- Suppose ~~$n_i Y_i \sim$~~ Bin $n_i Y_i \sim$ ~~Bin~~ Bin (n_i, Y_i), Y_i is the sample proportion of successes.

$$\ell(\beta) = \sum_{i=1}^{n} L_i = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi).$$

$\ell(\beta)$ – depends on β through θ_i

Likelihood eq. $\frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial L_i}{\partial \beta_j} = 0 \;\forall j.$

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j}$$

Since $\frac{\partial L_i}{\partial \theta_i} = [y_i - b'(\theta_i)]/a(\phi)$, $\mu_i = b'(\theta_i)$

$$\frac{\partial L_i}{\partial \theta_i} = \frac{y_i - \mu_i}{a(\phi)}, \quad \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \text{var}(y_i)/a(\phi)$$

$\eta_i = \sum \beta_j x_{ij}$, $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$

and finally $\eta_i = g(\mu_i)$, $\frac{\partial \mu_i}{\partial \eta_i}$ depends on the link function

$$\frac{\partial L_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a(\phi)} \cdot \frac{a(\phi)}{\text{var}(y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$$

$$= \frac{(y_i - \mu_i) x_{ij}}{\text{var}(y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i}$$

Summing over all obs.

$$\sum \frac{(y_i - \mu_i) x_{ij}}{\text{var}(y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 0, 1, 2, \ldots$$

$$\mu_i = g^{-1}(\sum \beta_j x_{ij})$$

- The LR equations depend on the dist of Y_i only through μ_i and $\text{var}(Y_i)$
- The variance depends on the mean through a particular form

$\text{var}(Y_i) = V(\mu_i)$ for some variance fn $V(\cdot)$

For example $V(\mu_i) = \mu_i$, for Poisson.

$V(\mu_i) = \mu_i(1-\mu_i)/n_i$ for Binomial

$V(\mu_i) = \sigma^2$, a constant for normal

Likelihood eq for Binomial GLMs

Suppose $n_i Y_i \sim \text{Bin}(n_i, \pi_i)$ (Y_i is the sample proportion of successes for n_i trials)

$E(Y_i) = \pi_i$

The Binomial GLM's for single predictor, $[\pi(x) = \Phi(\alpha+\beta x)]$ extends with several predictors to.

$\pi_i = \Phi(\sum \beta_j x_{ij})$, Φ - standard cdf of some cont - dist

$\sum \frac{(y_i - \pi_i)x_{ij}}{\pi_i(1-\pi_i)/n_i}$

$\pi_i = \mu_i = \Phi(\eta_i)$

$\frac{\partial \mu_i}{\partial \eta_i} = \phi(\eta_i) = \phi(\sum \beta_j x_{ij})$

$\phi(u) = \frac{d\Phi(u)}{du}$

Since $\text{Var}(Y_i) = \pi_i(1-\pi_i)/n_i$, the likelihood eq simplifies to.

$\sum \frac{n_i (y_i - \pi_i)x_{ij}}{\pi_i(1-\pi_i)} \phi(\sum \beta_j x_{ij}) = 0$, $\pi_i = \Phi(\sum \beta_j x_{ij})$

$j = 0, 1, \ldots$

For the logit link,

$\Rightarrow \sum \eta_i (y_i - \mu_i) x_{ij} = 0 \quad , j = 0, 1, \ldots, p.$

Likelihood eq for Poisson log link model

$Y_i \sim Poi(\mu_i)$

$E(y_i) = \mu_i \qquad Var(y_i) = \mu_i$

$$\sum \frac{(y_i - \mu_i) x_{ij}}{Var(y_i)} \frac{\partial \mu_i}{\partial \eta_i}$$

$$\Rightarrow \sum_i \frac{(y_i - \mu_i) x_{ij}}{\mu_i} \mu_i = 0 \Rightarrow \sum_i (y_i - \mu_i) x_{ij} = 0$$

$$\eta_i = \log \mu_i = \sum \beta_j x_{ij}$$

$$\Rightarrow \frac{\partial \eta_i}{\partial \mu_i} = 1/\mu_i \Rightarrow \frac{\partial \mu_i}{\partial \eta_i} = \mu_i$$

Asymptotic Covariance Matrix of Model Parameter Estimates

- The likelihood fn for the GLM also determines the asymptotic covariance matrix of the ML estimator $\hat{\beta}$.
- Information matrix $I = \left(E\left[\frac{\partial^2 l}{\partial \beta_i \partial \beta_j} \right] \right)$

  L_i: contribution from the ith sample fit to the log likelihood.
- For the exponential family, we have

$$E\left[\frac{\partial^2 L_i}{\partial \beta_i \partial \beta_j} \right] = -E\left(\frac{\partial L_i}{\partial \beta_i} \right) \left(\frac{\partial L_i}{\partial \beta_j} \right)$$

$$E\left[\frac{\partial^2 l(\beta)}{\partial \beta_k \partial \beta_j}\right] = \sum_{i=1}^{n} \frac{x_{ik} x_{ij}}{Var(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

- Let W be a diagonal matrix with ith diagonal entry

$$w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 / Var(y_i)$$

Then $I = X^T W X$

↳ the form of W and hence I depend on the link function.

- The asymptotic variance-covariance matrix of $\hat{\beta}$ is estimated by $\hat{Cov}(\hat{\beta}) = \hat{I}^{-1} = (X^T \hat{W} X)^{-1}$

where $\hat{W}$ is W evaluated at $\hat{\beta}$

In case of Poisson log linear model, the estimated asymptotic covariance matrix will be

$$w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 / Var(y_i) = \mu_i$$

$$\log \mu_i = \sum_j \beta_j x_{ij}$$

Deviance & Goodness of fit

- The standard GLM has a seperate parameter for each observation. It gives a perfect fit.
- Sounds good, but not a helpful model.
- Does not smooth data and hence no advantage of having a simpler model.
- It serves a baseline for other models, for example, for checking model fit.

Let, $\tilde{\theta}$ denote the estimate of $\underline{\theta}$, for the saturated model, corresponding is estimated means $\tilde{\mu}_i = y_i \ \forall i$

- For a particular unsaturated model, denote the corresponding ML estimators by $\hat{\underline{\theta}}$ and $\hat{\mu}_i$.
- $l(\hat{\underline{\mu}}, \underline{y})$: maximized log likelihood for the unsaturated model.
- $l(\underline{y}; \underline{Y})$: " " "

saturated model.

- $-2\left[l(\hat{\underline{\mu}}; \underline{Y}) - l(\underline{y}; \underline{Y}) \right]$ describes the lack of fit

- LR statistic for testing the null hypothesis that the model holds against the alt fact a more general model holds.

$$-2\left[l(\hat{\mu}; \underline{y}) - l(\gamma; \underline{y})\right]$$

$$= 2\sum \left[y_i \tilde{\theta}_i - b(\tilde{\theta}_i)\right] / a(\phi) - 2\sum \left[y_i \hat{\theta}_i - b(\hat{\theta}_i)\right] / a(\phi).$$

When $a(\phi) = \phi/\omega_i$ this is equal to

$$= 2\sum_{i=1}^{n} \omega_i \left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\right] / \phi$$

$$= \frac{D(\gamma; \hat{\mu})}{\phi}$$

is called the scaled deviance & $D(\underline{y}; \hat{\mu})$ is the deviance.

Deviance for Poisson GLMs

$\hat{\theta}_i = \log \hat{\mu}_i$ — ll f^n saturated model

$b(\hat{\theta}_i) = \exp(\hat{\theta}_i) = \hat{\mu}$

$\tilde{\theta}_i = \log y_i$

$b(\tilde{\theta}_i) = y_i$ ~~$\theta_i = \log x_i$~~

. $a(\phi) = 1$, so the deviance and scaled deviance are equal and

$$D(y; \hat{\mu}) = 2\sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\mu_i}\right) - y_i + \hat{\mu}_i\right]$$

- When a model with log link contain an intercept term

The likelihood eq of Poisson GLM with log link

$$\sum (y_i - \mu_i) x_{ij} = 0$$

$\sum (y_i - \mu_i) > 0$ corresponding to intercept x_{ij}'s

$$\sum y_i = \sum \hat{\mu}_i$$

$$D(\underline{y}; \hat{\mu}) = 2 \sum \left[y_i \log (y_i / \hat{\mu}_i) \right]$$

$a(\phi) = 1/n_i \Rightarrow \phi = 1$, $w_i = n_i$ [Binomial GLM].

Likelihood Ratio Model Comparison using deviances

for a poisson / binomial model denoted by μ, $\phi = 1$, So

$$D(\underline{y}; \underline{\hat{\mu}}) = -2 \left[l(\hat{\mu}, \underline{y}) - l(\underline{y}, \underline{y}) \right]$$

- Consider two models M_0 with fitted values $\hat{\mu}_0$ & μ_1 with fitted values $\hat{\mu}_1$, with μ_0 a special case of μ

Since μ_0 is simpler than μ, a smaller set of parameter values satisfy μ_0 than satisfy μ,

- maximizing the log-likelihood over a smaller space cannot yield a larger maxima

$$l(\hat{\mu}_0; \underline{y}) \leq l(\hat{\mu}; \underline{y})$$

- With the same model for each μ_0 & μ_1 $[l(\underline{y}; \underline{y})]$

$$D(\underline{y}; \hat{\mu}_1) \leq D(\underline{y}; \hat{\mu}_0)$$

→ simpler models have larger deviances.

- Assuming that model μ_1 holds, the LR test of hypothesis that μ_0 holds uses the test statistic.

$$= -2\left[l(\hat{\mu}_0, \underline{y}) - l(\hat{\underline{\mu}}; \underline{y})\right]$$

$$= -2\left[l(\hat{\mu}_0, y) - l(y; \underline{y})\right] - \{-2\left[l(\hat{\mu}_1; Y) - l(y; \underline{y})\right]$$

$$= D(\underline{y}; \hat{\mu}_0) - D(y; \hat{\mu}_1)$$

- The LR statistic comparing two models is simply the difference between the deviance.

- This is large when μ_0 fits poorly compared to μ_1

$$D(y; \hat{\underline{\mu}}_0) - D(\underline{y}; \hat{\mu}_1) =$$

$\hat{\mu}_0 \to \hat{\theta}_{0i}$
$\hat{\mu}_1 \to \hat{\theta}_{1i}$

$$= 2 \sum_i w_i \left[y_i(\hat{\theta}_{1i} - \hat{\theta}_{2i}) - b(\hat{\theta}_{1i}) + b(\hat{\theta}_{0i})\right]$$

approx χ^2_{df} df: difference between the no of parameters in the two models.

under certain regularity conditions.

for Poisson ~~log likelihood~~ log linear model with intercept.

$$D(\underline{y}; \hat{\underline{\mu}}) = 2\sum_i y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right)$$

$$D(\underline{y}; \hat{\mu}_0) - D(\underline{y}; \hat{\mu}_1) = 2\sum_i y_i \log\left(\hat{\mu}_{1i}/\hat{\mu}_{0i}\right)$$

NR: $\hat{\underline{\beta}}^{(t+1)} = \hat{\beta}^{(t)} = (H^{(t)})^{-1} \underline{u}^{(t)}$

assuming that the Hessian matrix $H^{(t)}$ is ns.

Fisher scoring $\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + \left(I^{(t)}\right)^{-1} u^{(t)}$

$I^{(t)}$: information matrix.

(β is ~~measured~~ replaced by $\beta^{(t)}$).

$u^{(t)}$: sets of first order derivative evaluated at $\hat{\beta}^{(t)}$

$I = X' W X$, W is diagonal with

$$w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 / V(Y_i)$$

$I^{(t)} = X' W^{(t)} X$, $W^{(t)}$ evaluated at $\hat{\beta}^{(t)}$.

- for both Fisher scoring & MR algorithm, the score j^{th} u has the form

$$u_j = \frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{V(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}$$

Define $D = \text{diag}\{d_1, \ldots, d_n\}$, $d_i = \frac{\partial \mu_i}{\partial \eta_i}$

The GLMs likelihood eqn can be written as

$$u = X' W D^{-1} (Y - \mu) = 0$$

- Consider maximizes the loglikelihood based on obs y from $\text{Bin}(n, \pi)$ dist^n.

$$l(\pi) = y \log \pi + (n-y) \log (1-\pi)$$ [Take π as the unknown parameter]

$$u = (y - n\pi) / \pi(1-\pi) \; ;$$

$$H = -\left[\frac{y}{\pi^2} + \frac{n-y}{(1-\pi)^2}\right]$$

Each NR step has the form

$$\pi^{(t+1)} = \pi^{(t)} + \left[\frac{y}{(\pi^{(t)})^2} + \frac{n-y}{(1-\pi^{(t)})^2}\right]^{-1} \frac{y - n\pi^{(t)}}{\pi^{(t)}(1-\pi^{(t)})} \quad (*)$$

(i) in positive when

$y/n > \pi^{(t)}$

$\pi^{(t+1)}$ increases if $y/n > \pi^{(t)}$ & decreases when $y/n < \pi^{(t)}$.

When $\pi^{(0)} = 1/2$; $\pi^{(1)} = \frac{y}{n}$

with $\pi^{(t)} = y/n$; no adjustment occurs & $\pi^{(t+1)} = y/n$ which is the correct answer.

- for starting values other than 0.5 ; adequate convergence usually happens in just few more steps.
- The information is $\frac{n}{\pi(1-\pi)}$ in this case.
- so the steps for fisher scoring stages

$$\pi^{(t+1)} = \pi^{(t)} + \left[\frac{n}{\pi^{(t)} + (1-\pi^{(t)})}\right]^{-1} \frac{y - n\pi^{(t)}}{\pi^{(t)}(1-\pi^{(t)})}$$

$$= \pi^{(t)} + \frac{y - n\pi^{(t)}}{n} = y/n$$

This gives the correct answer for $\hat{\pi}$ after a single criterion and stays at that value for for successive iteration.

ML as Iterative Reweighted least Square

$$\underline{Z} = X\underline{\beta} + \underline{\epsilon} \quad ; \quad V(\underline{\epsilon}) = V$$

The WLS of $\underline{\beta}$ is $(X'V^{-1}X)^{-1} X'V^{-1}\underline{Z}$

$$T = X'WX$$

From fisher scoring

$$I^{(t)}\,\hat{\underline{\beta}}^{(t+1)} = I^{(t)}\,\hat{\underline{\beta}}^{(t)} + \underline{u}^{(t)}$$

$$= (X'W^{(t)}X)\,\hat{\underline{\beta}}^{(t)} + X'W'(D^{(t)})^{-1}(\underline{y} - \underline{\mu}^{(t)})$$

$$= X^{T}W^{(t)}\left[X\hat{\underline{\beta}}^{(t)} + D^{(t)^{-1}}(\underline{y} - \underline{\mu}^{(t)})\right]$$

$$= X'W^{(t)}\underline{z}^{(t)}$$

When $\underline{z}^{(t)}$ has element

$$z_i^{(t)} = \sum_j x_{ij}\beta_j^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}}$$

$$= \eta_i^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}}$$

$$I^{(t)}\,\hat{\beta}^{(t+1)} = I^{(t)}\,\hat{\beta}^{(t)} + \underline{u}^{(t)}$$

$$\Rightarrow (X'W^{(t)}X)\,\hat{\underline{\beta}}^{(t+1)} = X'W^{(t)}z^{(t)}$$

$$\hat{\underline{\beta}}^{(t+1)} = (X^{T}W^{(t)}X)^{-1}X^{T}W^{(t)}\underline{z}^{(t)}$$

- The vector $\underline{z}^{(t)}$ is an estimated linearized form of the link f^{n}, evaluated at y_i.

$$g(y_i) \approx g(\mu_i^{(t)}) + (y_i - \mu_i^{(t)})\,g'(\mu_i^{(t)})$$

$$= \eta_i^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}} = z_i^{(t)}$$

· At cycle t: regresses $\underline{z}^{(t)}$ on x with weight $w^{(t)}$

- obtain new estimate $\hat{\beta}^{(t+1)}$
- new linear perdictor $\underline{\eta}^{(t+1)} = X\hat{\beta}^{(t+1)}$
- new adjusted $\underline{z}^{(t+1)}$

· The ML estimator results from iterative use of WLS in which the weight matrix changes at each cycle.

→ It is called iterative reweighted LS.

$\mu_i^{(0)} = y_i$, Bernoulli response ; $y_i \in \{0,1\}$.

logit

Poisson response , $\mu_i = y_i > 0$

$y_i \geq 0$, we start $y_i = 1$

$y_i \geq 0, 1$ $\quad$ $y_i \geq 1/2$

Residuals for GLM

· Pearson residual for obs i

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\mu_i)}} \quad ; v(y_i) = v(\mu_i)$$

Deviance residual:

$$D(y;\hat{\mu}) = 2\sum_i w_i \left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right]$$

$$= \sum d_i$$

$\tilde{\theta}_i$ → saturated

$\hat{\theta}_i$ →

Deviance residual is $\sqrt{d_i}\ \text{sign}(y_i - \hat{\mu}_i)$

↳ sum of sq is the deviance.

Logistic Regression:

Interpretating parameters in logistic Regression

· for binary response variable y and an expl variable x, let, $\pi(x) = P(y=1 \mid x=x) = 1 - P(y=0 \mid x=x)$

The logistic regression model

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

or logit has the linear relationship.

$$\text{logit}[\pi(x)] = \log \frac{\pi(x)}{1-\pi(x)} = \alpha+\beta x \quad \text{—(2)}$$

Q: How can we interpret β.

- Its sign determines whether $\pi(x)$ ↑ or ↓ as x increases.

· The rate of climb or descent increases on $|\beta|$ increases as $\beta \to 0$, the curve flatters to a horizontal st. line.

· When $\beta = 0$; y is independent of x

· for quantitative x with $\beta > 0$, the curve for $\pi(x)$ has the slope of the CDF of the logistic dist^n.

· Some logistic density is symmetric, $\pi(x)$, approaches 1 at the same rate is it approaches to zero.

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

↳ The odds are exponential f^n of x

The odds multiply by e^{β} for 1 unit increase in x.

- In other words, e^{β} is an odds ratio.

↳ odds at $x = x+1$ / odds at $x = x$

- Since it has a curved rather than a linear appearance, the logistic regression f^n (1)

⇒ the rate of change in $\pi(x)$ per unit change in x varries

$$\frac{\partial \pi(x)}{\partial x} = \frac{\beta e^{\alpha+\beta x}}{(1+e^{\alpha+\beta x})^2}$$

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

The line tangent to the curve at x for which $\pi(x) = 1/2$ has slope. $\beta/4$

- ~~Slope $\longrightarrow 0$ as $\pi(x)$~~
- When $\pi(x) = 0.9$ or 0.1 ; slope $= 0.09\beta$
- Slope $\to 0$ as $\pi(x) \to 1.0$ or 0

The steepest slope occurs at x for which $\pi(x) = 1/2$

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\alpha+\beta x) \Rightarrow x = -\alpha/\beta.$$

- This x values $(-\alpha/\beta)$ is sometimes called median effective level.

In taxilogy it is called lettal dose.

[the dose with 50% chance of a lethal result].

END

Categorial dat analysis

→ Book (GLM for)