The LR statistic

$$-2\left[LR(\hat{\beta},\hat{\sigma}^2) - l(\hat{\beta},\hat{\lambda})\right] \left[-2\log[\text{LR statistic}]\right]$$

$$\sim \chi^2_K \quad \text{(asymptotically under } H_0\text{)}$$

Score Test:

- Has advantange that we donot have to calculating the unrestricted MLEs.
- For a general wt function w,

$$\frac{\partial l}{\partial \underline{\lambda}} = -\frac{1}{2}\sum_{i=1}^{n}\left\{\frac{1}{w_i} - \frac{e_i^2}{w_i^2}\right\}\frac{\partial w_i}{\partial \underline{\lambda}}$$

Let, D: the matrix whose ith row is is $\frac{\partial w_i}{\partial \underline{\lambda}}$

$\underline{u} = (u_1, \ldots, u_n)'$; $u_i = \frac{e_i^2}{\hat{\sigma}^2}$

& $\hat{w}_i = \hat{\sigma}^2$ under the null

$$\left.\frac{\partial l}{\partial \underline{\lambda}}\right|_{H_0} = -\frac{1}{2}\sum_{i=1}^{n}\frac{\hat{\sigma}^2 - e_i^2}{\hat{\sigma}^4}\cdot\frac{\partial w_i}{\partial \underline{\lambda}}$$

Differentiate and taking expectation

$$I_{H_0} = \begin{bmatrix} I_{\beta\beta} & \underline{0} \\ \underline{0} & I_{\underline{\lambda}\underline{\lambda}} \end{bmatrix}$$

$$\frac{\partial l}{\partial \underline{\beta}} = X'\Sigma^{-1}(\underline{y} - X\underline{\beta}) \quad ; \quad \frac{\partial^2 l}{\partial \beta \partial \beta'} = -(X'\Sigma^{-1}X)$$

$$[\Sigma = \sigma^2 I] \qquad = -\sigma^{-2}(X'X) \quad \text{under } H_0$$

$$\frac{\partial^2 l}{\partial \lambda \partial \underline{\beta}'} = \frac{\partial}{\partial \underline{\beta}}\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \underline{x}_i'\underline{\beta})^2}{w_i^2}\frac{\partial w_i}{\partial \underline{\lambda}}$$

(first term is independent of $\underline{\beta}$)

$$= \sum_i \frac{(y_i - \underline{x}_i'\underline{\beta})}{\omega_i^2} \frac{\partial \omega_i}{\partial \underline{\lambda}} \underline{x}_i'$$

$\longrightarrow$ zero expectation.

$$\frac{\partial^2 l}{\partial \underline{\lambda} \partial \underline{\lambda}'} = -\frac{1}{2} \sum_{i=1}^{n} \left(\frac{1}{\omega_i} - \frac{\epsilon_i^2}{\omega_i^2} \right) \frac{\partial^2 \omega_i}{\partial \underline{\lambda} \partial \underline{\lambda}'} + \frac{1}{2} \sum_{i=1}^{n} \left(\frac{1}{\omega_i^2} - \frac{2\epsilon_i^2}{\omega_i^3} \right) \frac{\partial \omega_i}{\partial \underline{\lambda}} \left(\frac{\partial \omega}{\partial \underline{\lambda}} \right)'$$

$$E\left[\frac{1}{\omega_i} - \frac{\epsilon_i^2}{\omega_i^2} \right] = 0$$

$$E(\epsilon_i^2) = \omega_i$$

$$E\left[\frac{1}{\omega_i^2} - \frac{2\epsilon_i^2}{\omega_i^3} \right] = -\sigma^{-4} \quad [\text{under the null}]$$

$$= \frac{1}{\omega_i^2} - \frac{2\omega_i}{\omega_i^3} = -\frac{1}{\omega_i^2}$$

$$= -\sigma^{-4}$$

$$E\left(\frac{\partial^2 l}{\partial \underline{\lambda} \partial \underline{\lambda}'} \right) = -\frac{1}{2} D'D \sigma^{-4}$$

$$\mathcal{I}_{H_0} = \begin{pmatrix} \sigma^{-2}(X'X) & 0 \\ 0 & \frac{1}{2}\sigma^{-4}(D'D) \end{pmatrix}$$

The score statistic

$$\left(\frac{\partial l}{\partial \underline{\lambda}} \right)' \mathcal{I}_{\lambda\lambda}^{-1} \left(\frac{\partial l}{\partial \underline{\lambda}} \Big|_{H_0} \right) \sim \chi_m^2$$

[m: no of restrictions imposed by the null.]

So the score statistic takes the form

$$\frac{1}{2} (\underline{u} - \underline{1}n)' D (D'D)^{-1} D' (\underline{u} - \underline{1}n)$$

If $\omega(z, \lambda) = \exp(z'\lambda)$, $D = \hat{\sigma}^2 Z$ under the null.

& the test statistic is

- If variances are known f^n of unknown parameter

$$\sigma_i^2 = \omega(z_i, \lambda, \beta) \quad \text{——} \circledast\circledast$$

↳ This is to allow for the possibility that the variance depends on the mean and possibly on other parameters.

$$\sigma_i^2 = \lambda_1 (x_i'\beta)^{\lambda_2} \quad ; \lambda = (\lambda_1, \lambda_2)'$$

Algorithm:

(I) obtain an estimate β^*, for example, by LS method.

(III) Compute $\hat{\beta}$ as $(X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}Y$

$$\hat{\Sigma} = \text{diag}\{\omega(z_1, \lambda^*, \beta^*), \ldots, \omega(z_n, \lambda^*, \beta^*)\}$$

using a WLS program.

(II) m th $\hat{\beta}$ fixed at β^*, obtain an estimate λ^* of λ

Regressing Residuals:

$e_i \approx \epsilon_i$, then $E(e_i^2)$ y, $\text{Var}(\epsilon_i) = \omega(z_i, \lambda, \beta^*)$

Obtain an estimate of λ by non-linear LS problem

$$\min_{\lambda} \sum_{i=1}^{n} \left[e_i^2 - \omega(z_i, \lambda, \beta^*) \right]^2$$

Since $\text{Var}(e_i^2) \approx \text{Var}(\epsilon_i^2) = 2\,\omega(\underline{z}_i \underline{\lambda}, \beta^*)$,

an alternative method is to solve

$$\min_{\underline{\lambda}} \sum_{i=1}^{n} \frac{\left(e_i^2 - \omega(\underline{z}_i, \underline{\lambda}, \underline{\beta}^*)\right)^2}{\omega(\underline{z}_i, \underline{\lambda}, \beta^*)}$$

Variance is a function of the mean:

- If the variance is a smooth, f^n of the mean, we can do better, Suppose that, $\sigma_i^2 = \omega(\underline{x}_i', \underline{\beta})$:

ω - Known

Algorithm: I: Obtain an estimate of $\underline{\beta}$, say $\hat{\underline{\beta}}_{OLS}$, the

II: Calculate $\hat{\Sigma} = \text{diag}\{\omega(\underline{x}_1, \underline{\beta}), \ldots, \omega(\underline{x}_n, \underline{\beta})\}$

III: Recompute $\hat{\underline{\beta}} = (X'\hat{\Sigma}^{-1}X)^{-1} X'\hat{\Sigma}^{-1}\underline{y}$

Note: (IV): Repeate II & III until conversence.

Carrol & Repperl showed that the estimate provided by this algorithm has the same asymptotic efficiency as $\hat{\underline{\beta}}_{WLS}$

- This means, in this case, there is no cost, in not knowing the weights provided ω is known

- The same is true if ω is unknown but smooth.

Carroll showed if we plot the squared residuals from a LS fit v/s the OLS fitted values & smooth the plot, we can do just as well.

Let, $\hat{\omega}_i$ be the smoothed value of e_i^2

Then if we use WLS with weights to $1/\hat{\omega}_i$, we

obtain an estimate whose asymptotic efficiency relative to $\underline{\hat{\beta}}_{WLS}$ is 100 %.

obtain an estimate whose asymptotic efficiency relative to $\hat{\beta}_{WLS}$ is 100%.

Transforming to equalize variances:

We want to find an ~~increasing~~ increasing function f s.t. $f(y_i) = x_i'\beta + \epsilon_i$

where ϵ_i have equal variance

Suppose $\text{Var}(y_i) = \omega(\mu_i)$; $\mu_i = x_i'\beta$

Assuming ω is known.

$$\text{Var}[f(y)] \approx \left(\frac{df}{d\mu}\right)^2 V(y)$$

$f(y) \approx f(\mu) + (y-\mu) f'(\mu)$

The variance of the transformed responses $f(y_i)$ will be approximately constant if we choose f so that

$$[f'(\mu)]^2 \omega(\mu) \text{ is constant}$$

$$f(\mu) := \int \frac{d\mu}{\{\omega(\mu)\}^{1/2}}$$

$E[f(y)] \approx f(\mu) + 0$

$\text{Var}[f(y)] \approx E[f(y) - f(\mu)]^2$

~~$\approx V(y)[f'(\mu)]^2$~~

$= E[(y-\mu) f'(\mu)]^2$

$= V(y)[f'(\mu)]^2$

Example: $y_i \sim \text{Poisson}(\mu_i)$; $\omega(\mu) = \mu_i$

$$f(\mu) = \mu^{1/2}$$

we expect $y_1^{1/2}, \ldots, y_n^{1/2}$ have approximately equal variance.

Example: $y_i \sim \text{Binomial}(m_i, p_i)$; $\mu = m_i p_i$

$$\omega(\mu) = \mu\left(1 - \frac{\mu}{m}\right)$$

Then, $$\int \frac{d\mu}{\sqrt{\mu(1-\mu/m)}} \propto \sin^{-1}\left(\left(\frac{\mu}{m}\right)^{1/2}\right)$$

If w_i is not known, we can experiment by the responses Y_i using a power transformation.

We can transform with a sample power less than 1 and use the diagonastic plots to see if the variances have been made equal.

- We proceed by reducing the power until the wedge effect in the plot of squared residuals v/s fitted values disapper.

Box-cox Transformation / Power family of Transformation

- A flexible family of transformation
- When the values of the response variable are all positive

$$g(y;\lambda) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & ; \lambda \neq 0 \\ \log_e y & ; \lambda = 0 \end{cases}$$

Power transformation in y^λ when $\lambda = 0$; $y^\lambda = 1$; making all data pt equal.

Departure from Normality:

- Non-normal errors are detected by normal plot of residuals.

(I) first the LS residuals.

(II) Plot these argument quantiles of the standard normal dist^n.

- Suppose the normality assumptions is satisfied.

$\underline{e}$ has a singular dist^n $N(\underline{0}, \sigma^2(I-H))$

$N(0, \sigma^2(I-H))$ is approximately $N(0, \sigma^2 I)$ provided h_{ij}'s are small.

Here residuals are approximately a random sample from a $N(0, \sigma^2)$ dist^n.

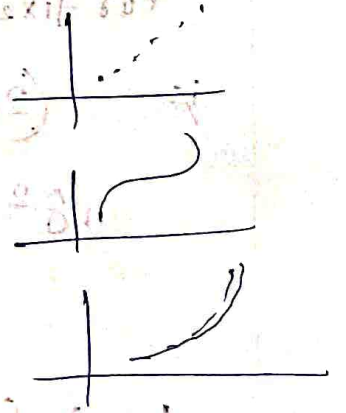
Normal prob. plot is a special case of Q-Q plot.

So that, $E[e_{(i)}] \approx \sigma z(\alpha_i)$

$\alpha_i = \frac{i - 0.5}{n}$; $z(\alpha)$ is the α th quantile of the standard normal distribution

$$\int_{-\infty}^{z(\alpha)} f(z)\,dz = \alpha$$

- Normal error - roughly a straight line.
- Skewed error - a curve.
- heavy-tailed dist^n - S-shaped.
- Outliers - isolated pts



Transforming the response

If the normal plot reveals non-normality, the standard remedy is to transform the response.

- Box-cox.
- Box-Cox introduced this transformation to remedy several type of regression problems, so that all the regression assumptions would be satisfied.

(Liv in exp variable, homogeneous variances, normal error).

Assume that there is a transformation parameter λ ?

$$y_i^{(\lambda)} = g(y_i; \lambda) = \underline{x}_i' \underline{\beta} + \underline{\epsilon}$$

Under this transformation the likelihood f^n for the original observations y_i's is

$$(2\pi\sigma^2)^{-n/2} \exp\left[-\tfrac{1}{2}\left(\underline{y}^{(\lambda)} - X\underline{\beta}\right)'\left(\underline{y}^{(\lambda)} - X\underline{\beta}\right)\right] |J| \quad (*)$$

for each $y_i > 0$

Here, $|J| = \left| \prod_{i=1}^{n} \frac{dy_i^{(\lambda)}}{dy_i} \right| = \prod_{i=1}^{n} y_i^{\lambda-1}$

For fixed λ, (*) is the likelihood corresponding to a standard LS problem except for the constant factor J.

For fixed λ, the maxm value of the likelihood f^n is $(2\pi\hat{\sigma}^2)^{-n/2} e^{-n/2} |J|$

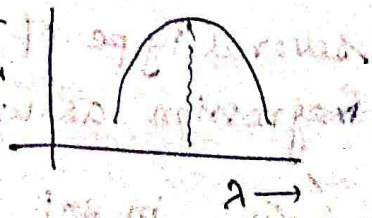
$$n\hat{\sigma}^2 = \underline{y}^{(\lambda)} \left(I - X(X'X)^{-1}X'\right) \underline{y}^{(\lambda)}$$

$$= RSS(\lambda; Y), \text{ say}$$

Apart from a constant, the maxm log likelihood is

$$l_{max}(\lambda) = -\tfrac{1}{2} \ln\{RSS(\lambda; Y)\} + (\lambda - 1) \sum_{i=1}^{n} \log y_i$$

Box and Cox suggested plotting $l_{max}(\lambda)$ against λ function for a trial series of values & take that λ as $\hat{\lambda}$ which maximizes $l_{max}(\lambda)$.



A more accurate value of $\hat{\lambda}$ can be obtained by solving $\frac{d\, l_{max}(\lambda)}{d\lambda} = 0$

Another family :

$$g(y;\lambda) = \begin{cases} sign(y)\ \{|y|+1\}^{\lambda}/\lambda & , \lambda \neq 0 \\ sign(y)\ \log\{|y|+1\} & , \lambda = 0 \end{cases}$$

- defined for all values of y.
- For each λ, g is a 1-1 monotone transformation
- general works well with symmetrically distributed observation with heavy tails.

Box-Cox is better in case of skewed data

↳ can behave very badly in presence of outliers

$$y_i = x_i'\beta + \epsilon_i$$

Transforming both Sides

If $E[y_i] = x_i'\beta$, but the errors are non-normal and/or heteroscedastic.

- Then transforming the response will destroy the linear from of the mean.
- To avoid this; transform both sides,

$$g(y_i;\lambda) = g(x_i'\beta, \lambda) + \epsilon_i$$

Where for some transformation $g(y;\lambda)$ and some value of λ, the errors ϵ_i are normally distributed with constant variance.

When $g(y;\lambda)$ is box-cox family, use ML method.

Then the transformed log likelihood -

$$l\left(\underline{\beta}, \sigma^2, \lambda\right) = c - \tfrac{1}{2}\left\{ n \log \sigma^2 + \frac{1}{\sigma^2} \sum_{i=1}^{n} \left\{ g(y_i : \lambda) - g(\underline{x}_i', \underline{\beta}) \right\}^2 \right\} + (\lambda - 1) \sum_{i=1}^{n} \log y_i$$

$$\frac{\partial l}{\partial \sigma^2} = 0 \Rightarrow \hat{\sigma}^2(\underline{\beta}, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(g(y_i; \lambda) - g(x_i' \underline{\beta}, \lambda) \right)^2$$

$$l_{max} = \left(\underline{\beta}, \hat{\sigma}^2, \lambda\right) = c - \tfrac{1}{2} n \log \hat{\sigma}^2(\beta, \lambda) + n/2 + (\lambda - 1) \sum_{i=1}^{n} \log y_i$$

↳ can be maximized by fisher Scoring to obtain the estimate of $\underline{\beta}$ & λ.

Alternatively, take $\left(\prod_{i=1}^{n} y_i \right)^{1/n} = \dot{y}$

Then the log likelihood upto a constant then

$$-n/2 \left[\log \hat{\sigma}^2 (\underline{\beta}, \lambda) - 2\lambda \log \dot{y} \right]$$

$$= -\frac{n}{2} \log \left[\frac{\hat{\sigma}^2 (\underline{\beta}, \lambda)}{(\dot{y})^{2\lambda}} \right]$$

→ is maximized by minimising

$$\sum_{i=1}^{n} \left[\frac{g(\beta, \lambda) - g(\underline{x}' \beta, \lambda)}{(\dot{y})^{\lambda}} \right]^2$$

- Standard error can calculated from the information matrix.
- This technique can be used to make error dist^n closer to normal and also stabilize the variance.
- It is not necessary true that the same

transformation will achieve both the objections.

- If the transformation that makes the error dist^n closer to normal does not also make the variance more homogeneous, it may be necessary to use weighting to achieve the constant variance (approx.) case.

Detecting Outliers and Dealing with Outliers.

$y_i = \underline{x}_i'\beta + \epsilon_i$; $\hat{y}_i = \underline{x}_i'\hat{\underline{\beta}} + e_i$

$\hat{\underline{\beta}}$ estimates $\underline{\beta}$.

$\hat{y}_i = x_i'\hat{\underline{\beta}}$; $e_i = y_i - \hat{y}_i \Rightarrow y_i = \hat{y}_i + e_i$

$= x_i'\hat{\beta} + e_i$

It is reasonable to consider e_i is an estimate of ϵ_i.

We can use the raw residuals e_i or the standardized residuals. r_i & ϵ_i

↳ identically distributed.

- Standard graphical plots such as box plot of residuals.
  ↳ we can identify an outliers, the points having large residuals.

- $t_i \sim t_{n-p-1}$; in absence of outliers, a reasonable a definition of "large" is a pt for which $|t_i| > 2$.

→ Th, diagnstic approach works well which does not have high leaverage.

- If it does we cannot expect the corresponding residual to reveal the presence of an outliers.

- Suppose that the ith-response is as $y_i + \Delta_i$ rather than y_i ; so, $x_i'\beta + \Delta_i + \epsilon_i$

Let, $\Delta = (0, 0, \ldots, \Delta_i, 0, \ldots, 0)'$

$$\underline{e} = (I - H)\underline{Y} \;;$$

$$E[\underline{e}] = (I - H)(X\beta + \Delta)$$

$$= (I - H)\Delta$$

$$E[e_i] = (1 - h_i)\Delta_i$$

If x_i is close to $\bar{x}$, h_i is small, we can expect the residual to reveal the outlier quite well.

⟶ Here $E[e_i]$ is else to Δ_i

→ If the data point is an high-leverage point, then hat matrix diagonal is to be close to 1, so the residual will be much smaller than Δ_i

- Hat matrix diagonals are reasonable measures of leaverage, can be interpreted in terms of MDi

- the average of hat matrix diagonal p/n, an arbitrary but reasonable definition of high-leaverage pt is one which satisfy

$$h_i > 2p/n$$

whether a point has high leaverage depends on the exp variables, included in the regression.

Q: Having identified that a pt is having high leaverage, how can you detect whether the pt is an outliers.

(1) A the effect that a data pt has on the regression by deleting and refitting the regression.

- If the regression quantity is change drastically, then the pt in is a high influence pt or an outlier.

↳ leave-one-out-diagonistic | regression coefficients, fitted values, standard errors, etc.

(2) Use a robust fitting method that is not affected by high-leaverage pts., resulting in residuals, that better identify the outlier.

Leave-one-out Diagonostics

Th: $\hat{\beta} - \hat{\beta}(i) = \frac{(X'X)^{-1} x_i e_i}{1-h_i}$; x_i: ith row of X.

Change in Estimated regression coefficients

$\hat{\beta}(i)$: LSE of $\hat{\beta}$ when ith obs is deleted.

$\hat{\beta} - \hat{\beta}(i) = \frac{(X'X)^{-1} x_i' e_i}{1-h_i}$: the change is propotional to the size of the residual but is inflated of h_i is close to 1.

DFBETA:

Let, $C = (X'X)^{-1} X'$ → catcher matrix

The (j,i) th element of C is

$$c_{j,i} = \left[(X'X)^{-1} \underline{x}_i\right]_j$$

So, the jth element of DFBETA is

$$\frac{e_{j,i}\, e_i}{1-h_{ii}}$$

$$\sum_i e_{j,i}^2 = \sum_i \left[(X'X)^{-1} \underline{x}_i\right]_j^2$$

$$= \sum_i \left[(X'X)^{-1} \underline{x}_i \underline{x}_i' (X'X)^{-1}\right]_{jj}$$

$$= \left[(X'X)^{-1} \sum_i \underline{x}_i x_i' (X'X)^{-1}\right]_{jj}$$

$$= \left[(X'X)^{-1} X'X (X'X)^{-1}\right]_{jj}$$

$$= \left[(X'X)^{-1}\right]_{jj}$$

$$\left[CC' = (X'X)^{-1} X'X (X'X)^{-1} = (X'X)^{-1}\right]$$

- Standarize the $(j+1)$ th element of DFBFTA by an estimate of the standard error of $\hat{\beta}_j$, $s(i)\left[(X'X)^{-1}\right]^{1/2}_{j+1,\,j+1}$

$$= s(i)\left(\sum e_{j+1,i}^2\right)^{1/2}$$

$$\text{DFBETA}\, S_{i,j} = \frac{\hat{\beta}_j - \hat{\beta}(i)_j}{s(i)\left(\sum_i c_{j+1,i}^2\right)^{1/2}}$$

In terms of externally Studentized residuals t_i,

$$DFBETA\ S_{i,j} = \frac{c_{j+1,i}\, e_i}{s(i)\left(\sum_i c_{j+1,i}\right)^{1/2}(1-h_i)}$$

$$= \frac{c_{j+1,1} \cdot t_i}{\left(\sum_i c_{j+1,i}^2\right)^{1/2}(1-h_i)^{1/2}}$$

$$\left[\text{as } t_i = \frac{e_i}{s(i)\sqrt{1-h_i}}\right]$$

If the ith data pt is not an outlier and does not have high leaverage we expect that

$$|t_i| < 2.$$

and y_i will not have a large effect on the value of $\hat{\beta}_j$. Since, $\hat{\beta}_j = \sum_{i=1}^{n} c_{j+1,i}\, y_i$

The $c_{j+1,i}$ is small in relation to $\left(\sum c_{j+1}^2, i\right)^{-1/2}$

and approximately $n^{-1/2}$

A suitable cut off value for detecting large residuals using DFBETA $S_{i,j}$ is $2/\sqrt{n}$.

$\beta_1, \ldots, \beta_4$, $i = 1, 2, \ldots n$

$n \longrightarrow \hat{\beta}_1$

$n \longrightarrow \hat{\beta}_2$

DFFIT

Change in fitted values:

The change in the ith obs fitted values in

$$x_i' \hat{\beta} - x_i' \hat{\beta}(i) = \frac{x_i'(X'X)^{-1}e_i}{1-h_i}$$

↳ DFFIT.

$$= \frac{h_i e_i}{1-h_i}$$

Standarize by the estimated standard error.

$s(i)\, h_i^{1/2}$

$$\text{DFFITSS}_i = \frac{h_i e_i}{s(i)\, h_i^{1/2}(1-h_i)}$$

$$= \frac{h_i^{1/2} e_i}{s(i)(1-h_i)}$$

$$= t_i \left(\frac{h_i}{1-h_i}\right)^{1/2}$$

$\text{var}(x_i'\hat{\beta}) = h_i\sigma^2$
$= x_i' V(\hat{\beta}) x_i$
$= \sigma^2 x_i'(X'X)^{-1}x_i$
$= \sigma^2 h_i$

For pts, which are not outliers and do not have high leverage, $|t_i| < 2$ & h_i will not be far away from the average value p/n.

The cutoff pt DFFITSS$_i$ is $2\sqrt{\frac{p}{n-p}}$ or even $2\sqrt{p/n}$.

Cook's Distance:

(Cook's D):

[Distance between $\hat{\beta}(i)$ & $\hat{\beta}$]

$$Z_1 = \frac{(\hat{\beta}-\beta)\, X'X\, (\hat{\beta}-\beta)}{\sigma^2} \sim \chi^2_p$$

$$Z_2 = \frac{RSS}{\sigma^2} = \frac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p}$$

$$\frac{Z_1/p}{Z_2/(n-p)} = \frac{(\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta)}{ps^2} \sim F_{p,n-p}$$

[$\hat{\beta}$ and s^2 are independent]

- So a $100(1-\alpha)\%$ Confidence ellipsoid for $\underline{\beta}$ is

$$\left\{\beta: (\underline{\beta}-\hat{\underline{\beta}})' X'X (\beta-\hat{\underline{\beta}}) \leq ps^2 F^{\alpha}_{p,n-p}\right\}$$

- Cook's distance measure D_i is defined as the distance of $\hat{\underline{\beta}}$ and $\hat{\underline{\beta}}(i)$

$$D_i = \frac{(\hat{\underline{\beta}}(i) - \hat{\underline{\beta}})' X'X (\hat{\underline{\beta}}(i) - \hat{\underline{\beta}})}{ps^2} \quad 1 \leq i \leq n$$

- based on confidence ellipsoid.

$$D_i = \frac{x_i'(X'X)^{-1}(X'X)(X'X)^{-1}x_i e_i^2}{(1-h_i)^2 ps^2}$$

$$= \frac{e_i^2}{(1-h_i)^2} \cdot \frac{h_i}{ps^2} = \frac{r_i^2 h_i}{p(1-h_i)}$$

pts with large D_i have considerable influence on the LSE of β.

Cook's suggestion $D_i > F^{0.10}_{p,n-p}$

- Similar to Sq of DFFITSS (differing only in use of r_i instead of t_i and constant divisor p).

Another way of writing Cook's D

$$D_i = \frac{(\hat{\underline{y}}(i) - \hat{\underline{y}})'(\hat{\underline{y}}(i) - \hat{\underline{y}})}{ps^2}, \quad X\hat{\beta}(i) = \hat{y}_{(i)}$$

Covariance Ratio:

The estimated variance-covariance matrix of $\hat{\beta}$ is $s^2 (X'X)^{-1}$ & its leave-one-out version is $s(i)^2 [X(i)' X(i)]^{-1}$

$$COVRATIO_i = \frac{\det [s(i)^2 [X(i)' X(i)]^{-1}]}{\det [s^2 (X'X)^{-1}]}$$

$$(n-p-1)\, s(i)^2 = (n-p)\, s^2(i) - \frac{e_i^2}{1-h_i} \quad ; \left[s(i)^2 = \frac{e_i^2}{1-h_i}\right]$$

$$\Rightarrow \frac{s^2}{s(i)^2} = \frac{n-p-1}{n-p} + \frac{e_i^2}{s(i)^2 (1-h_i)(n-p)}$$

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

$$\det [X(i)' X(i)] = \det [X'X - x_i x_i']$$

$$= |X'X| \left[1 - x_i' (X'X)^{-1} x_i\right] \quad \left[\det(A + uu') = |A|\,|1 + u' A^{-1} u|\right]$$

$$= |X'X| [1 - h_i]$$

$$\text{So, } COVRATIO_i = \frac{|s(i)^2 \{X(i)' X(i)\}^{-1}|}{|s^2 (X'X)^{-1}|}$$

$$= \left(\frac{s(i)^2}{s^2}\right)^p \frac{|(X'(i) X(i))^{-1}|}{|(X'X)^{-1}|}$$

$$= \left(\frac{n-p-1}{n-p} + \frac{t_i^2}{n-p}\right)^{-p} \frac{1}{1-h_i}$$

Extreme Cases:

① $|t_i| > 2$, but this case has the minimum leverage with $h_i = 1/n$ [$h_i \geq 1/n$ for regression models with intercept term]

$$COVRATIO_i = \left(1 + \frac{t_i^2 - 1}{n-p}\right)^{-p} \frac{n}{(n-1)}$$

$$\approx \left[1 - \frac{(t_i^2 - 1)p}{n-p}\right] \cdot \frac{n}{n-1}$$

$$\approx 1 - \frac{(t_i^2 - 1)p}{n}$$

when $n >> p$. Therefore if $|t_i| > 2$

$$COVRATIO_i \approx 1 - \frac{(t_i^2 - 1)p}{n} \leq 1 - 3p/n$$

② The opposite situation in when t_i is small but (in absolute value) the case have high leverage.

Take $t_i = 0$, $h_i > 2p/n$

$$COVRATIO_i = \left(1 - \frac{t_i^2}{n-p}\right)^{-p} (1-h_i)^{-1}$$

$$\geq \left(1 - \frac{1}{n-p}\right)^{-p} \left(1 - \frac{2p}{n}\right)^{-1}$$

$$\approx \left(1 + \frac{p}{n-p}\right)\left(1 + \frac{2p}{n}\right)$$

$$\approx \left(1 + p/n\right)\left(1 + \frac{2p}{n}\right) \text{ for large } n.$$

$$\approx 1 + \frac{3p}{n}$$ [ignoring the higher order terms]

The cases having $\left|COVRATIO_i - 1\right| > \frac{3p}{n}$ are considered to have high influence.

[for high lev pt, COVRATIO will large]

Deleting Several Influential points

Consider there are $m (>1)$ obs to be assessed for influenced simultaneously.

Let, i denote the $m \times 1$ vector of indices specifying the pts to be assessed.

Define,

$$D_i = \frac{(\hat{\beta}(i) - \hat{\beta}) \times \times (\hat{\beta}(i) - \hat{\beta})}{p s^2}$$

→ extension of single obs Cook's D.

Large value of D_i indicates that the set of m pts are influential.

- Selection of the set of pts to include in i, is not obvious, a in some situations, the m pts are jointly influential but are not so.
- It is also ~~different~~ difficult to choose all m-subsets of the n pts co-investigate

Leave-many-out

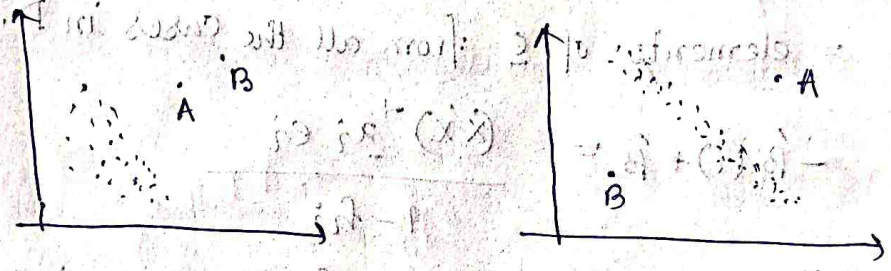
Single point deletion diagonistic will not be able to identify a pt as an outlier if it is masked by other.

→ leave-d-out diagonistic

Drawbacks

Practically applicable for small values of d

It will be effective in identifying the situation only if all the points in a cluster are deleted.



If the cluster contains d pts, we must calculate leave-d-out diagonistics

- This is a problem since d is not known in advance
- The procedure is computationally ~~infeasible~~ feasible for small values of d.
- Let D be the d subset of cases.

$X(D)$: $(n-d) \times p$ submatrix of X corresponding to the rows of the cases not in D.

X_D : $d \times p$ submatrix corresponding to the rows of the cases in D.

Then $\left[X(D)' X(D)\right]^{-1} = \left[X'X - X_D' X_D\right]^{-1}$

$$\left[X(D)' X(D)\right]^{-1} = \left[X'X - X_D' X_D'\right]^{-1}$$

$$= (X'X)^{-1} - (X'X)^{-1} X_D' \left(Id - H_D\right)^{-1} X_D (X'X)^{-1}$$

$$\left[\begin{array}{l} A_{m\times m},\ B_{n\times n} \longrightarrow ns \quad U_{m\times n},\ V_{n\times n} \\ (A+UBV')^{-1} = A^{-1} - A^{-1}UB\left(B - BVA^{-1}UB\right)^{-1} BVA^{-1} \end{array}\right]$$

$$H_D = X_D (X'X)^{-1} X_D'$$

$$\underset{p\times 1}{\hat{\beta}_{(D)}} = \hat{\beta} - (X'X)^{-1} X_D' (I_d - H_D)^{-1} e_D$$

e_D = elements of e from all the cases in D.

$$-\hat{\beta}(i) + \hat{\beta} = \frac{(X'X)^{-1} x_i e_i}{1 - h_i}$$

The difference in fitted values for the cases in D.

$$\hat{y}_D - \hat{y}_D(D) = X_D(\hat{\beta}) - X_D \hat{\beta}(D)$$

$$= H_D (I_D - H_D)^{-1} e_D$$

The analog of the COVRATIO can be shown

$$\left\{ \left[\frac{n-p-d}{n-p} + \frac{e_D' (I_d - H_D)^{-1} e_D}{n-p} \right] \det(I_d - H_D) \right\}^{-1}$$

Test for outliers:

To test of a fixed set of K observations contains outliers, assuming, the remaining $n-k$ observations are "clean"

Arrange the data so that the clean obs^n come first, followed by the possibly outlying obs.

Outlier Stuff Model:

$$Y = X\beta + Z\gamma + \epsilon$$

$$Z = \begin{pmatrix} 0 \\ I_K \end{pmatrix} \begin{matrix} \leftarrow n-k \times k \\ \leftarrow K \times k \end{matrix}$$

γ : a K-vector containing the shifts for the outlying observation.

To test $\underline{\gamma} = \underline{0}$

Let, H be the hat matrix for the regression

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

Let, $H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}$; $H_{11} \to (n-k) \times (n-k)$

Partitioning the residual $(I_n - H)\underline{Y} = \underline{e}$ as $\underline{e} = (\underline{e}_1', \underline{e}_2')'$

$\underline{e}_1 = (n-k) \times 1$

Then the LSE of $\underline{\gamma}$ is

$$\hat{\underline{\gamma}} = (Z'(I-H)Z)^{-1} Z'(I-H)\underline{Y}$$

$$= \left[(0, I_k)(I_n - H)\begin{pmatrix} 0 \\ I_k \end{pmatrix}\right]^{-1} (0, I_k)\begin{pmatrix} \underline{e}_1 \\ \underline{e}_2 \end{pmatrix}$$

$$= (I_k - H_{22})^{-1}\underline{e}_2$$

$$F = \frac{(RSS_H - RSS)/q}{RSS/(n-p)} \quad ; \quad RSS_H = \underline{Y}'(I-H)\underline{Y}$$

$$RSS = \underline{Y}' R_G \underline{Y} = \underline{Y}' R \underline{Y}$$

$$= \hat{\underline{\gamma}}_G Z' R \underline{Y}$$

$$\hat{\underline{\gamma}}' Z' R \underline{Y} = \hat{\underline{\gamma}}'(0, I_k)(I-H)\underline{Y}$$

$$R = I - H.$$

$$= \hat{\underline{\gamma}}' \underline{e}_2$$

$$= \underline{e}_2' (I_k - H_{22})^{-1} \underline{e}_2$$

So, $RSS_H - RSS = \underline{e}_2' (I_k - H_{22})^{-1} \underline{e}_2$

So, the F-test for $\underline{\gamma} = 0$ is

$$F = \frac{[\underline{e}_2' (I_k - H_{22})^{-1} \underline{e}_2]/k}{[RSS - \underline{e}_2' (I_k - H_{22})^{-1} \underline{e}_2]/(n-p-k)} \sim F_{k, n-p-k}$$

RSS → RSS of model $\underline{Y} = X\underline{\beta} + \underline{\epsilon}$

When $k=1$, $\underline{e}_2' (I_k - H_{22})^{-1} \underline{e}_2 = \frac{e_i^2}{1-h_i}$

(corresponding to ith obs).

$$F = \frac{(n-p-1)\, e_i^2}{(1-h_i)\, RSS - e_i^2} \sim F_{1,n-p-1} \quad (\text{under } H_0)$$

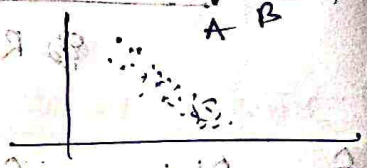
(when ith obs is not an outlier)

Using $r_i^2 = \frac{e_i^2}{(1-h_i)\, s^2}$

$$F = \frac{(n-p-1)\, r_i^2}{\left(n-p-r_i^2\right)}$$

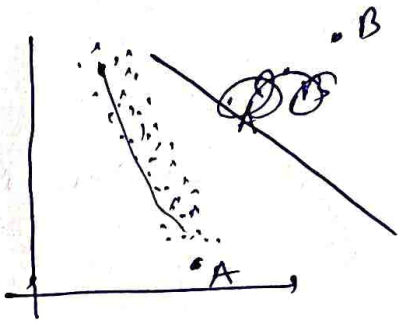
Masking and Swamping:

- Hat matrix diagonal and leave-one-out diagonistic are useful tools, but sometimes cannot identify outliers & high-leverage pt.



- Taken singly, points A and B are not influential, since either will have little effect on the fitted line, because remaining pts tend to affect the line.
- Taken as a pair, they are influential. (we say that)
- In this case, pts A & B mask each other.
- To use robust method in the least median squares to identify the cluster or use leave many out (diagonostics).

Pt A will have large residuals, since B is attracting the line away from A.

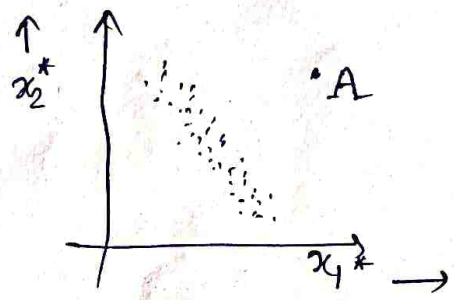


Swamping: Whom pts that are not outlying can be mistaken as outliers.

- Sampling does not occur when robust methods are used.

p=3, two centered & scaled explanatory variables:

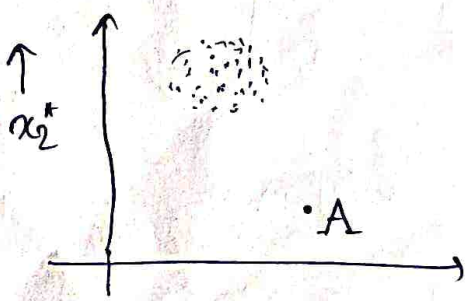
x_1^*, x_2^*



K(-i)
→ after removing ith obsⁿ, then calculate & condⁿ no.

- Compare K and K(i).

[K: CN on all obsⁿ].



Basic Time Series:

Defⁿ: (The Auto covariance function (ACVF)).

If $\{X_t, t \in \mathbb{Z}\}$ is a process $\ni \operatorname{Var}(X_t) < \infty \ \forall t \in \mathbb{Z}$

then the ACVF of $\{X_t\}$ is

$$\gamma_X(r,s) = \operatorname{Cov}(X_r, X_s)$$
$$= E\left[X_r - E(X_r)\right]\left[X_s - E(X_s)\right],$$
$$r, s \in \mathbb{Z}$$

Stationary:

The TS $\{X_t : t \in \mathbb{Z}\}$ with index set $\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$ is said to be (weak) stationary if

(i) $E|X_t|^2 < \infty \quad \forall t \in \mathbb{Z}$

(ii) $E[X_t] = m \quad \forall t \in \mathbb{Z}$

(iii) $\gamma_X(r,s) = \gamma_X(r+t, s+t) \quad \forall r, s, t \in \mathbb{Z}$

This is also known as Covariance stationary, stationary in wide sence or second order stationary.

If $\{X_t ; t \in \mathbb{Z}\}$ is stationary $\gamma_X(r,s) = \gamma_X(r-s, 0)$

$\forall r, s \in \mathbb{Z}$

So we redefine the ACVF of a stationary TS as

$$\gamma_X(h) = \gamma_X(h,0) = \operatorname{Cov}(X_{t+h}, X_t), \ t, h \in \mathbb{Z}$$

ACF (Auto Correlation Function) of $\{X_t\}$ is

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \operatorname{Cov}(X_{t+h}, X_t) \quad \forall t, h \in \mathbb{Z}$$

Defⁿ: The TS $\{X_t ; t \in \mathbb{Z}\}$ is strictly stationary of the joint distⁿ of $(X_{t_1}, X_{t_2}, \ldots X_{t_k})'$ and $(X_{t_1+h}, X_{t_2+h}, \ldots X_{t_k+h})'$ and the sence for all positive integer k and for all $t_1, t_2 \ldots, t_k, h \in \mathbb{Z}$

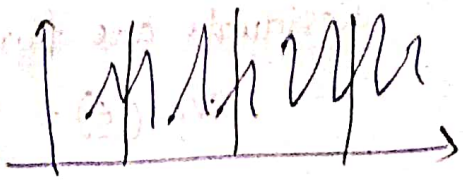
Proposition:

If $\gamma(\cdot)$ is the ACVF of a stationary process, then

$\gamma(0) \geq 0$

$|\gamma(h)| \leq \gamma(0) \quad \forall h \in \mathbb{Z}$

$\gamma(h) = \gamma(-h), \quad h \in \mathbb{Z}$



MA(q): [Moving Average of order q]

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q \cdot Z_{t-q};$$

$Z_t \sim WN(0, \sigma^2)$

uncorrelated

AR(p): [Autoregressive of Order p]

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \cdots - \phi_p X_{t-p} = Z_t;$$

$Z_t \sim WN(0, \sigma^2)$

Z_t is uncorrelated with X_s, $s < t$

ARMA(p,q)

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-1} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$$

MA(1):

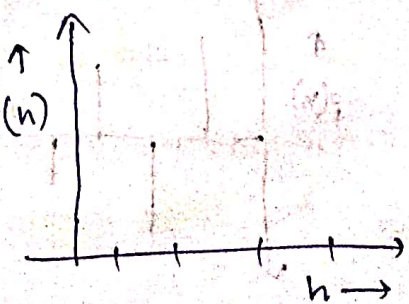
$$X_t = \theta Z_{t-1} + Z_t$$

$$V(X_t) = \gamma(0) = \sigma^2(1+\theta^2)$$

$$\text{Cov}(X_t, X_{t+1}) = \gamma(1) = \theta\sigma^2$$

$$\text{Cov}(X_t, X_{t+h}) = \gamma(h) = 0 \quad , \forall h \geq 2$$

$$\rho(1) = \theta/(1+\theta^2) = \frac{\gamma(1)}{\gamma(0)}$$



Residuals are from an MA(1) process,

$$\text{Var}(\underline{\epsilon}) = \Sigma = \sigma^2 \begin{pmatrix} 1+\theta^2 & \theta & 0 & \cdots & 0 \\ \theta & 1+\theta^2 & \theta & 0 & \cdots \\ \vdots & \theta & \ddots & & \\ 0 & \cdots & & \theta & \\ & & & -\theta & 1+\theta^2 \end{pmatrix}$$

$$\underline{Y} = X\underline{\beta} + \underline{\epsilon} \qquad ; \underline{\epsilon} \sim (0, \Sigma)$$

$\Sigma_{(1)}$ from the residual by estimating θ.

$$\hat{\beta}_{(1)} = \left(X' \Sigma_{(1)}^{-1} X\right)^{-1} X' \Sigma_{(1)}^{-1} \underline{Y}$$

Calculate the residual

Then θ $\Sigma_{(2)}$ & $\hat{\beta}_{(2)}$

$\hookrightarrow \hat{\underline{\beta}}, \hat{\theta}_1 \rightarrow \hat{Z}_t \;,\; \epsilon_t = Z_t + \theta Z_{t-1}$

$\hat{\epsilon}_t \rightarrow$

AR(1): $X_t = \phi X_{t-1} + Z_t \;,\; |\phi| < 1$

$$\gamma(h) = \text{Cov}(X_t, X_{t+h})$$

$$= \sigma^2 \sum_{j=0}^{\infty} \phi^j \phi^{j+h}$$

$$= \frac{\sigma^2 \phi^h}{1-\phi^2} \qquad ; h \geq 0$$

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}$$

$$= \phi^h \qquad ; h \geq 0$$

$$\gamma(0) = V(X_t) = \frac{\sigma^2}{1-\phi^2}$$

$(1-\phi B) X_t = Z_t$

$BX_t = X_{t-1}$

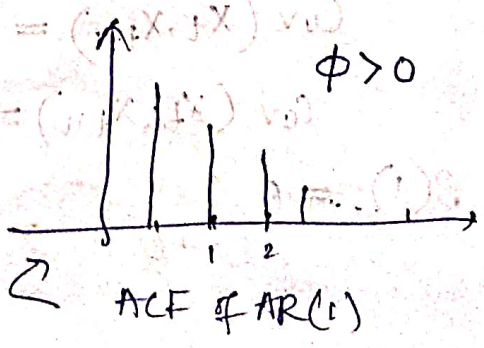
Backward shift operator

$X_t = (1-\phi B)^{-1} Z_t$

$= (1+\phi B + \phi^2 B^2 + \cdots) Z_t$

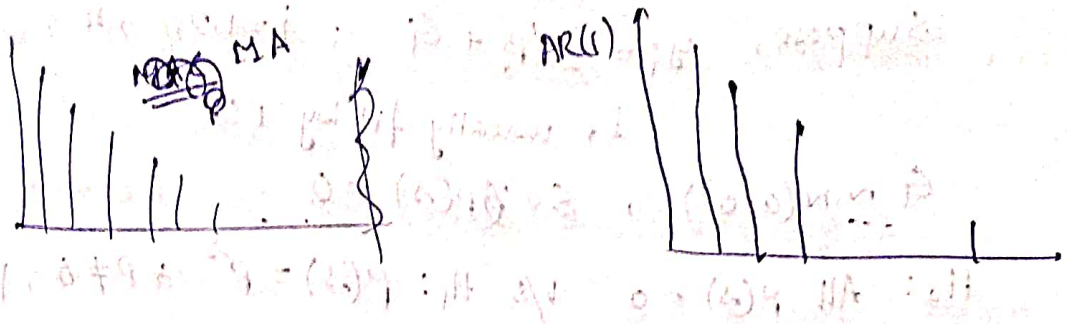
$= Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \cdots$

RACF (Partial Auto Correlation fn)



ACF of AR(1)

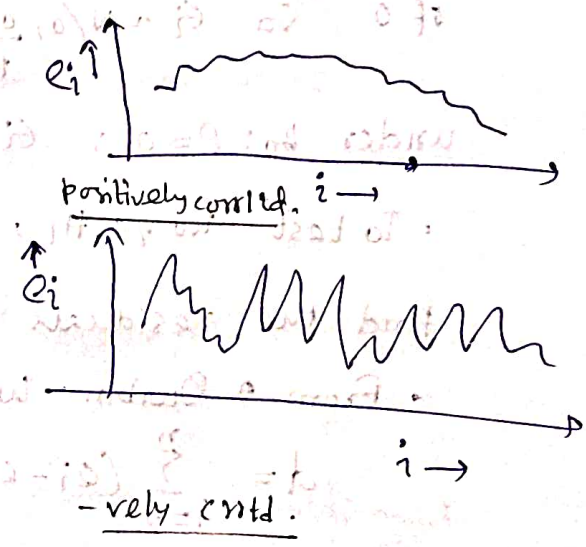
PACF (Partial Auto correlation function)



Serial Correlation and Durbin Watson Test:

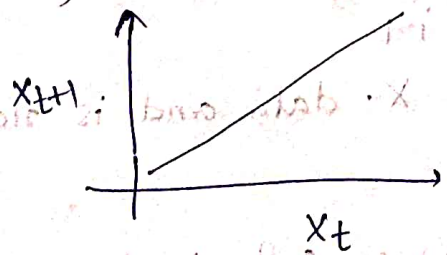
If the data are collected sequentially in time, then successive errors may be correlated.

If this is the case, a sequence plot of e_i against time order, ~~where~~ which is after e_i v/s i may show up the presence of any correlation



For positively correlated error, a residual tends to have the same sign as its predecessor, while for negatively correlated error, the signs.

(X_t, X_{t+1})



Divide the ~~data~~ time ordered residuals into consecutive pairs and plotting are member of the pair against the other.

Serially correlated data show up as a linear trend in the ~~Durbin~~ plot.

Durbian Watson Test :

Suppose, $y_i = \underline{x}_i' \underline{\beta} + \epsilon_i \quad ; \quad i = 1, 2, \ldots, n$

↳ usually fit by LS.

$\epsilon_i \sim N(0, \sigma^2)$, so $\rho(s) = 0$

H_0: All $\rho(s) = 0$ V/s H_1: $\rho(s) = \rho^s$; $\rho \neq 0$, $|\rho| < 1$

An alternative that arrizes from the assumption that the error ϵ_i are

$$\epsilon_i = \rho\, \epsilon_{i-1} + z_i \quad , \quad z_i \sim N(0, \sigma^2)$$

is indpt of $\epsilon_{i-1}, \epsilon_{i-2}, \ldots$

Also assume that mean & var of ϵ_i are, const, indpt of 0, so $\epsilon_i \sim N\left(0, \frac{\sigma^2}{1-\rho^2}\right)$.

under H_0: $\rho = 0$; $\epsilon_i \sim N(0, \sigma^2)$ ⟶ usual assumption

- To test H_0 V/s. H_1, fit the model $y_i = x_i' \underline{\beta} + \epsilon_i$ & find the residuals e_i.
- From ⊛ Durbin - Watson Statistic

$$d = \sum_{i=2}^{n} (e_i - e_{i-1})^2 \Big/ \sum_{i=1}^{n} e_i^2$$

- The dist^n of d depends on X - data and is not independent of them.
- The dist^n of d lies between 0 & 4 & is symmetric about 2.
- Percentage pts also depends on X-data & would have to be calculated each application to perform the test property.

- Test is usually performed using tabulated bounds (d_L, d_U) - instead of a single critical value, have to look into two critical values.
- d is used only for a lower tailed test $\rho > 0$.
- To test against $\rho < 0$, we need a upper tailed test - can be handled as a lower tailed test using the statistic 4-d.

1. One Sided Test: for $\rho > 0$.

If $d < d_L$, d is significant, reject H_0 at level α

If $d > d_U$, d is not " , do not reject H_0.

If $d_L \leq d \leq d_U$, the test is inconclusive.

2. One sided Test: $\rho < 0$

Repeat (1) using 4-d instead of d.

3. Two sided equal tailed test $\rho \neq 0$

If $d < d_L$ or $4-d < d_L$, d is significant, reject H_0 at level 2α.

If $d > d_U$ or $4-d > d_U$, d is not significant, do not reject H_0 at level 2α.

Otherwise, the test is inconclusive.

A simplified Test:

$H_0: \rho > 0$, If $d < d_L$, reject H_0 at level α, o.w. do not reject.

$H_0: \rho < 0$, If $4-d < d_U$, reject H_0 at level α, o.w. do not reject.

$H_0: \rho \neq 0$, If $d < d_L$ or $4-d > d_U$, reject H_0, at level 2α.

(D-W test).

Diagonosing Collinearity :

The square of the condition no is an upper bound for the VIFs and

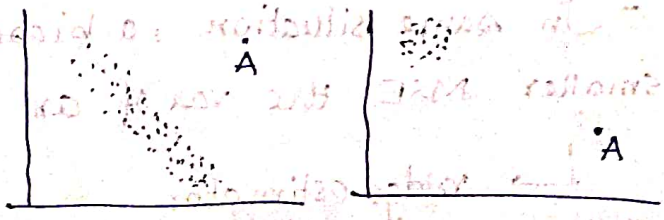
- If the condition no is small, there will be no estimated regression coefficients with large variance.
- If the cond^n no is large at least one eigen value must be small.
- The eigen value corresponding to the smallest eigen value should also be examined to determine the linear combinations that are causing the trouble.
- Can also calculated the variance proportions.
  - the proportion of each variance due to each eigen value.

The variance proportion for the jth variable eigen value λ_r.

The variance proportion for the jth var given value λ_r.

$$t_{jr}^2 \lambda_r^{-1} \Big/ \sum_i t_{ji}^2 \lambda_i^{-1}$$

- Proportion near unity indicate that most of the var of a particular regression coeffs is due to a single small eigen value.



- High leaverage pts can have a effect on collinearity.
- A more focused diagonostic is to compute the condition no $K(-i)$ of the regression with ith row removed.

- Comparison of $K(-i)$ and K (CN of the X matrix) will indicate any pts i.e. causing a drastic change.

Remedies of Collinearity:

① Collinearity arrises when the data are deficient.

↳ collect fresh data to repair the deficiencies in the reg matrix.

- not always possible, when the exp variable are strongly correlated in the popl^n from which the data were collected.

② Another way is simply to discard the variables until the data are not collinear.

- The deleted variables may have a relationship with the response.

③ Another way is to abandon the method of LS and use a biased estimation method, say ridge regression.

Ridge Regression:

- Introduced to deal with collinear data.
- LS# method might results in a very poor estimates of the regression coefficients which the data are collinear.
- LSE is unbiased.
- In same situation, a biased estimator may have a smaller MSE the var of an unbiased estimator.

↳ ridge estimator

Consider the scaled and centered reg. model

$$y_i = \alpha + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \cdots + \gamma_{p-1} x_{ip-1} + \epsilon_i$$

$$, i = 1, 2, \ldots, n.$$

The ridge estimate of $\underline{\gamma}$ is defined as

$$\left(X^{*\prime}X^{*} + kI\right)^{-1} X^{*\prime}\underline{y}$$

where k is a tve parameter whose value must be specified k is not necessarily an integer when $k=0$, ridge estimate reduces to LSE.

Interpolation: It is basically the LSE of the data will augmented.

- Imagine augmenting the observed data (X^{*}, y) with raw data $(kI_{p-1}, 0)$
- The LSE calculated from the augmented data

$$\begin{pmatrix} X^{*} & y \\ kI_{p-1} & 0 \end{pmatrix} \qquad \left[\left(X^{*\prime} \; kI_{p-1}\right)\begin{pmatrix} X^{*} \\ kI_{p-1} \end{pmatrix}\right]^{-1}\left(X^{*\prime} \; kI_{p-1}\right)\begin{pmatrix} \underline{y} \\ 0 \end{pmatrix}$$

is just the ridge estimate.

$$\begin{aligned}
\hat{\underline{\gamma}}[k] &= \left[X^{*\prime}X^{*} + kI_{p-1}\right]^{-1} \underline{X}^{*\prime}\underline{y} \\
&= \left[R_{xx} + kI_{p-1}\right]^{-1} R_{xx} R_{xx}^{-1} X^{*\prime} Y \\
&= \left[R_{xx} + kI_{p-1}\right]^{-1} R_{xx}\, \hat{\underline{\gamma}} \\
&= \left\{R_{xx}\left(I_{p-1} + kR_{xx}^{-1}\right)\right\}^{-1} R_{xx}\hat{\underline{\gamma}} \\
&= \left(I_{p-1} + kR_{xx}^{-1}\right)^{-1} \hat{\underline{\gamma}} \\
&= C_k\, \hat{\underline{\gamma}}
\end{aligned}$$

So, the ridge estimate is clearly biased as $C_k \neq I_{p-1}$

$$MSE = E\left[\|\hat{\underline{\gamma}}(k) - \underline{\gamma}\|^{2}\right] = E\left[\|C_k\hat{\underline{\gamma}} - \underline{\gamma}\|^{2}\right] = E\left[\|C_k\hat{\underline{\gamma}} - C_k\underline{\gamma} + C_k\underline{\gamma} - \underline{\gamma}\|^{2}\right]$$

$$= E\left[\|C_k(\hat{\underline{\gamma}} - \underline{\gamma})\|^{2}\right] +$$

$$= E\left[\| C_k(\hat{\gamma}-\gamma) + (C_k - I_{p-1})\gamma \|^2 \right]$$

$$= E\left[\| C_k(\hat{\gamma}-\gamma) \|^2 \right] + \| (C_k - I_{p-1})\gamma \|^2$$

$$\text{I} \qquad\qquad \text{II}$$

[Cross product is 0 as $E(\hat{\gamma}) = \gamma$]

$\hat{\gamma} - \gamma$ has mean zero and variance $\sigma^2 R_{xx}^{-1}$

$$\text{I} = E\left[\| C_k(\hat{\gamma}-\gamma) \|^2 \right] = E\left[(\hat{\gamma}-\gamma)' C_k' C_k (\hat{\gamma}-\gamma) \right]$$

$$= \sigma^2 \, tr(R_{xx}^{-1} C_k' C_k)$$

[Spectral decomposition of R_{xx} : $R_{xx} = T \Lambda T'$

$\Lambda = diag\{\lambda_1, \ldots, \lambda_{p-1}\}$ ↑ eivalues of R_{xx}.

$C_k = TDT'$

$D = diag\{ \frac{\lambda_i}{k+\lambda_j} \}$

$$= \sigma^2 tr\left[T \Lambda^{-1} T' \right]$$

$$= \sigma^2 tr\left[T\Lambda^{-1} T'TDT'TDT' \right)$$

$$= \sigma^2 tr\left[\cdot T \Lambda^{-1} D^2 T' \right]$$

$$= \sigma^2 tr\left[\Lambda^{-1} D^2 T'T \right]$$

$$= \sigma^2 tr\left[\Lambda^{-1} D^2 \right]$$

$$= \sigma^2 \sum_{j=1}^{p-1} \frac{\lambda_j}{(k+\lambda_j)^2}$$

Take $\alpha = T'\gamma$

$$\text{II} = \gamma' (C_k - I_{p-1})' (C_k - I_{p-1}) \gamma$$

$$= \alpha' T^{-1} \left[TDT' - I_{p-1} \right]' \left[TDT' - I_{p-1} \right] (T')^{-1} \alpha$$

$$= \underset{p-1}{\alpha'} (D - I_{p-1})' (D - I_{p-1}) \alpha$$

$$= \sum_{j=1} \alpha_j^2 k^2 / (k+\lambda_j)^2$$

So, $$MSE = \sum_{j=1}^{p-1} \frac{\alpha_j^2 k^2 + \sigma^2 \lambda_j}{(k+\lambda_j)^2}$$

The derivative w.r.t. k

$$\sum_{j=1}^{p-1} \frac{2\lambda_j (\alpha_j^2 k - \sigma^2)}{(k+\lambda_j)^3} \longrightarrow \text{negative for small}$$

values of h, so that for k sufficiently small, MSE decreases as k increases.

Solve $$\sum_{j=1}^{p-1} \frac{2\lambda_j (\alpha_j k - \sigma^2)}{(k+\lambda_j)^3} = 0$$

To find the value of k corresponding to the smallest MSE.

The minimising value depends on the unknowns $\underline{\alpha}, \sigma^2$.

↳ use $\hat{\alpha} = T'\hat{\gamma}$ & s^2

There is no guarantee that the result in $\hat{k}$ will lead to the smaller MSE than LSE.

But generally, it works well.

If centered & scaled exp variables are orthogonal, then $R_{xx} = I_{p-1}$ and eigen values are all 1.

Then $$2\sum \frac{\hat{\alpha}_j^2 k - s^2}{(k+1)^3} = 0$$

$$\Rightarrow \hat{k} = \frac{(p-1)s^2}{\sum_{j=1}^{p-1} \hat{\alpha}_j^2} = \frac{(p-1)s^2}{\|\hat{\gamma}\|^2}$$

It gives good results in non-orthogonal regression.

– much better than LSE.

Dealing with Curvature.

conditional mean of the response y given the exp variable $\rightarrow E(Y|X)$

We want to visualize the surface. whether ~~it can~~ $(M|\underline{x})$ can be adequately represented by a Mean of $\underline{x}$.

- Single x variable $\rightarrow$ a plot y v/s x.
  $\rightarrow$ smooth the plot.

If the realisation in linear - fit a linear model.
If not, try to transform y using a power transformation - or try to use a polynomial model.

Two Exp Variable:

Plot Y v/s $\underline{x} = (x_1, x_2)$ using a 3d plot.
- using dynamic rotation, the relationship may be revealed.

More than two expl variables:

- not possible to visualize the surface directly.
  A useful plot: residuals v/s fitted values.
  $\hookrightarrow$ are independent under normality assumption.
- If the linear model is correct, then the plot should ~~be~~ display a horizontal ~~pattern~~. band of pts.
- A curved regression surface reveals as a curved plot.
- smoothing will enhance the interpretation.

Disadvantage:

The nature of the curvature is not revealed.

More seriously it is possible for the regression surface to be curved without the e_i v/s $\hat{y}_i$ – plot revealing the curvature.

Partial residual Plot (PRP)

- Plots of suitably modified residuals v/s the exp variables

Suppose the true reg surface is

$$E(Y|\underset{\sim}{x}) = \beta_0 + \beta_1 g(x_1) + \underset{\sim}{\beta_2}' \underset{\sim}{x_2} \quad \text{---} \circledast \qquad \underset{\sim}{x} = (x_1, \underset{\sim}{x_2}')'$$

and g is unknown f^n.

$$E(Y|\underset{\sim}{x}) = \beta_0 + \beta_1 x_1 + \underset{\sim}{\beta_2}' \underset{\sim}{x_2}$$

PRP are designed to reveal the form of g.

Suppose we fit $\beta_0 + \beta_1 x_1 + \underset{\sim}{\beta_2}' \underset{\sim}{x_2}$

when $\circledast$ is the true model.

$$Y = \beta_0 + \beta_1 x_1 + \underset{\sim}{\beta_2}' \underset{\sim}{x_2} + \underbrace{[\beta_1 g(x_1) - \beta_1 x_1 + \epsilon]}_{\text{error point.}}$$

$$\approx \hat{\beta}_0 + \underset{\sim}{\hat{\beta}_1} x_1 + \underset{\sim}{\hat{\beta}_2}' \underset{\sim}{x_2} + e$$

Expect that the residual e from the linear fit would approximate $\beta[g(x_1) - x_1] \approx \hat{\beta}_1 [g(x_1) - x_1]$

A plot of $e_i^* = e_i + \hat{\beta} x_{i1} \approx \hat{\beta}_1 g(x_{i1})$ v/s x_{i1} might reveal the shape of g.

e_i^* : partial residuals.

Features:

A LS fit through the origin of this will be $\hat{\beta}_1$.

The residuals from this fit are original residuals

$$e_i^* = e_i - \hat{\beta}_1 x_{i1} = e_i + \hat{\beta}_1 x_{i1} - \hat{\beta}_1 x_{i1} = e_i$$

Drawbacks: (1) The appearance of the plot may over emphasize the importance of the exp. variable x_1 in the fit.

(2) Imp: If g is highly non-linear then the LSEs of β_2 are not very good estimates & this arguments break down.

Example: $p_3 = 3$, scaled and centered x

$$\hat{\gamma}_1 = \frac{1}{1-r^2}\left[\underline{x}^{*(1)\prime}\underline{y} - r\,\underline{x}^{*(2)\prime}\underline{y}\right] \qquad y = \alpha_0 + \gamma_1 x_1^* + \gamma_2 x_2^* + \epsilon$$

$$\underline{x}^{*(1)} = (x_{11}^*, x_{21}^*, \ldots, x_{n1}^*)', \quad \underline{x}_2^*$$

$$r = \underline{x}^{*(1)\prime}\underline{x}^{*(2)}$$

$$y = X_S \underline{\delta} + \epsilon$$

$$X_S = (\underline{1}, X^*)$$

$$\delta = (\alpha, \beta_1, \ldots \beta_{p-1})'$$

$$H = P = \frac{1}{n}\underline{1}\underline{1}' + \underline{x}^{*(2)}\underline{x}^{*(2)\prime} + \frac{1}{1-r^2}\left(\underline{x}^{*(1)} - r\underline{x}^{*(2)}\right)\left(\underline{x}^{*(1)} - r\underline{x}^{*(2)}\right)'$$

Now suppose that $\underline{Y} = \alpha \underline{1_n} + \gamma_1 \underline{g} + \gamma_2 x^{*(2)} + \underline{\epsilon}$

When $X_S = \left(1_n, x^{*(1)}, x^{*(2)}\right)$; $\underline{g}$

$\underline{g} = \left(g(x_{11}^*), \ldots, g(x_{n1}^*)\right)'$

So, $\left(I_n - H\right) X_S = \not{0}\ 0$

$$E[\underline{e}] = E\left[(I_n - H)\,\underline{Y}\right] = (I_n - H)\left(\alpha 1_n + \gamma_1 \underline{g} + \gamma_2 \underline{x}^{*(2)}\right)$$

$$= \gamma_1 (I_n - H)\,\underline{g}$$

$$E(\hat{\gamma}_1) = \frac{1}{1-r^2}\left[\underline{x}^{*(1)} - r\,\underline{x}^{*(2)}\right]' E(\underline{Y})$$

$$= \frac{1}{1-r^2}\left[\underline{x}^{*(1)} - r\,\underline{x}^{*(2)}\right]' \left(\alpha 1_n + \gamma_1 \underline{g} + \gamma_2 x^{*(2)}\right)$$

$$= \frac{1}{1-r^2}\,\gamma_1 \left(\underline{x}^{*(1)} - r\,\underline{x}^{*(2)}\right)' \underline{g} \qquad \left[\underline{x}^{*(j)\prime} \underline{1_n} = 0;\ j = 1,2\right.$$

$$\underline{x}^{*(j)\prime} \underline{x}^{*(j)} = 1$$

$$\left.\underline{x}^{*(1)\prime} \underline{x}^{*(2)} = r\right]$$

$$E[e_i^*] = E[\underline{e}] + \underline{x}^{*(1)} E[\hat{\gamma}_1]$$

$$= \gamma_1 (I_n - H)\,\underline{g} + \frac{\gamma_1}{1-r^2}\, x^{*(1)} \left[x^{*(1)} - r\,x^{*(2)}\right]' \underline{g}$$

$$= \gamma_1 \left(g - \bar{g}\,\underline{1_n}\right) - \frac{1}{1-r^2}\,\gamma_1\, \underline{x}^{*(2)} \left(\underline{x}^{*(2)} - r\,x^{*(1)}\right)' \underline{g}$$

→ indicates that the plot does not reveal the shape of g, but the shape $\bar{g} \rightarrow$ is contaminated by additional terms.

— The plot will suffer a large amount of contamination if the correlation between columns is large.

Adding and dealing variables :

Added variable plot: Assume that a constant term has been included in the model.

- Suppose that there is a ~~reago~~ reasonable st line relationship in the PRP for x_j
- The strength of the relationship is measured by the corrⁿ $\tilde{r}_j$ between x_{ij} and $e_{ij}^* = e_{ij} + \hat{\beta} x_{ij}$

$$\tilde{r}_j = \frac{\hat{\beta}_j^2 \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}{RSS + \hat{\beta}_j^2 \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}$$ [Show it]

$$VIF_j = \sigma^{-2} \, Var(\hat{\beta}_j) \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

VIF_j will be large if x_j can be predicted accurately from the other explanatory variables.

If F_j is the f-statistic for testing $H_0: \beta_j = 0$,

then $F_j = \frac{\hat{\beta}_j^2}{s^2 (x'x)^{-1}_{j+1, j+1}} = \frac{\hat{\beta}_j^2 \Sigma (x_{ij} - \bar{x}_j)^2}{s^2 VIF_j}$

Hence, $\tilde{r}_j^2 = \frac{F_j \cdot VIF_j}{(n-p) + F_j \cdot VIF_j}$; $RSS = (n-p) s^2$

- can get a better picture if the variable x_j ~~to the regression~~ should be added to the regression x_j and the residuals from a fit ~~ext~~ excluding the variables x_j.

If $X = (1_n, x^{(1)}, \ldots x^{(p-1)})$ & $x^{(j)}$ as X with $x^{(j)}$ omitted.

The $e^{(j)} = (I_n - P_j)\, y$ where $P_j = X^{(j)}(X^{(j)\prime} X^{(j)})^{-1} X^{(j)\prime}$

- When considering whether to add x_j to the regression, we must access how will we can predict the residuals $e^{(j)}$ by using x_j
- Since x_j can be partly predicted by the other exp. var we ~~can~~ are interested in how the residuals $e^{(j)}$ can be predicted by the pt of x_j that is not predicted by the other exp variable

- The suggest at the ~~variable~~ relationship between $e^{(j)}$ and x_j residuals $(I_n - P_j)\, x^{(j)}$

- The plot of $e^{(j)} = (I_n - P_j)\, y$ v/s $(I_n - P_j)\, x^{(j)}$

↳ A showing linear relationship $e^{(j)} \to x^{(j)}$ indicating that the variable be included $(I_n - P_j)\, x^{(j)}$ in the regression.

→ Added variable plot.

P_j^2 → Squared correlation between quantities in the added variable plot

$P_j^2 = \dfrac{F_j}{n - p + f_j}$; F_j → Statistic for testing $H_0: \beta_j = 0$

Model Selection:

To decide which variables to include in the model & how to construct a predictor form the variables available

- Assume that a large no K of potential exp var are available.
- Sometime, we assume that the true model is one of the form $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \epsilon$ —(1)
- The problem is to identify the variables that are not related to the response i.e. identifying the β's that are zero.

Two approaches

(I) All possible Regression (APR) (Require 2^K reg models)

- Define a criterion for goodness.
- Compute the criterion for each possible subset of variables.
- Choose the subset which optimizes the criterion.

The criterion can be based on

- Standard goodness of fit criterion.
- On estimating the prediction error.
- On estimating some measure of distance between the could based on subset & the true model.

(II) To apply a seq of hypothesis tests to the problem and attempt to identify the non-zero's Betas in (1).

- forward selection, backward elimination, stepwise regression.

- Assumption: that (1) is the true model

- Computationally less demanding, but there is no guarantee that model found will be optional in terms of criteria in (1)

- alternative to subject selection & LS is to use all exp variable & use a based estimation method that "shinks" the coefficients towards zero.

Prediction Error

- An initial data set $(y_i, \underline{x}_i)$, $i=1,2,\ldots,n$ containing of $n(p+1)$ dim multivariable observations.
- Assume that the first element of each $\underline{x}_i$ is 1, corresponding to the constant term
- Training set: we use this set of observation to predict responses y_{0i}, $i=1,2,\ldots,m$, corresponding to m new vector $\underline{x}_{0i}$; $i=1,2,\ldots,m$.

$$\underline{Y} = (y_1, \ldots, y_n)' \quad ; \quad Y_0 = (y_{01}, \ldots, y_{0m})'$$

- Assume that $V(\underline{y}) = \sigma^2 I_n$; $\text{Var}(Y_0) = \sigma^2 I_m$, $\underline{Y}_0$ & $\underline{Y}$ are independent.

$$X = \begin{pmatrix} \underline{x}_1' \\ \vdots \\ \underline{x}_n' \end{pmatrix}, \quad X_0 = \begin{pmatrix} \underline{x}_{01}' \\ \vdots \\ \underline{x}_{0m}' \end{pmatrix}$$

- Calculate $\hat{\underline{\beta}} = (X'X)^{-1}X'Y$ from the training set.
- The ISE of Y_0 (also the estimate of $\underline{\mu}_0 = E(Y_0)$) is $X_0\hat{\beta}$.

- The sum of squared prediction error is

$$\sum_{i=1}^{m} (Y_{0i} - x_{0i}'\hat{\beta})^2 = \| \underline{Y} - X_0 \hat{\beta} \|^2$$

$$= \| \underline{Y}_0 - \underline{\mu}_0 + \underline{\mu}_0 - X_0 \hat{\beta} \|^2$$

$$= \| \underline{Y}_0 - \underline{\mu}_0 \|^2 + \| \underline{\mu}_0 - X_0 \hat{\beta} \|^2 + 2 (\underline{Y}_1 - \underline{\mu}_0)' (\underline{\mu}_0 - X_0 \hat{\beta})$$

Take expectation over the new data only and call it the prediction error (PE).

$$PE = E_{Y_0} \| \underline{Y}_0 - X_0 \hat{\beta} \|^2$$

$$= E_{Y_0} \| \underline{Y}_0 - \underline{\mu}_0 \|^2 + \| \underline{\mu}_0 - X_0 \hat{\beta} \|^2 \quad \text{---} (*)$$

$$E_{Y_0} \left[(\underline{Y}_0 - \underline{\mu}_0)' (\underline{\mu}_0 - X_0 \hat{\beta}) \right] = 0$$

$$(*) \Rightarrow E_{Y_0} \left[\sum_{i=1}^{m} (Y_{0i} - \mu_{0i})^2 \right] + \| \underline{\mu}_0 - X_0 \hat{\beta} \|^2$$

$$= m\sigma^2 + \| \underline{\mu}_0 - X_0 \hat{\beta} \|^2$$

$$PE = m\sigma^2 + \| \underline{\mu}_0 - X_0 \hat{\beta} \|^2$$

↓ reflects the underlying variability of the data.

↳ measures how well the linear model represented by X_0 estimates the mean μ_0 of the new responses Y_0

- $\| \underline{\mu}_0 - X_0 \hat{\beta} \|^2 \rightarrow$ model error (ME).

- If $X_0 = X$, then $m = n$ & $\underline{\mu}_0 = E[\underline{Y}] = \underline{\mu}$, say

Then $\underline{\epsilon} = \underline{Y} - \underline{\mu}$; $P = X(X'X)^{-1}X'$.

$$ME = \| \underline{\mu} - X\hat{\beta} \|^2 = \| \underline{\mu} - P\underline{Y} \|^2$$

$$= \| \underline{\mu} - P(\underline{\mu} + \underline{\epsilon}) \|^2 = \| (I_n - P)\underline{\mu} - P\underline{\epsilon} \|^2$$

$= \| (I_n - P)\mu \|^2 + \| PE \|^2 = \mu'(I_n - P)\mu + \epsilon' P \epsilon$

[Cross product term consider as $P^2 = P$]

$E(ME) = \mu'(I-P)\mu + \sigma^2 tr(P)$

$= \mu'(I-P)\mu + \sigma^2 p.$

The corresponding

$E(PE) = E[n\sigma^2 + ME] = (n+p)\sigma^2 + \mu'(I_n - P)\mu$

Define total bias & total variance of the prediction $X\hat{\beta}$:

Total Bias $= \| \mu - E(X\hat{\beta}) \|$

Total Variance $= tr(V(X\hat{\beta})) = \sigma^2 p$

$\therefore E(ME) = \mu'(I-P)\mu + \sigma^2 P$

$\Rightarrow$ sq. of the total bias which variables of $\mu \in C(X)$; $\sigma^2 P$ → total variance

- As new variables are added to the model, the variability increases, but the bias decreases; unless the new variables are linearly dependent on the old variables already included in the model.
- The best model, having the smallest expected ME (& hence the smallest expected PE) will be some compromise between these two conflicting requirements of small bias and small variability.

Suppose $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$

is exactly the true model, although some of the reg coeff may be equal to zero.

$\underset{n\times p}{Y} = \underset{n\times(k+1)}{X}\underset{(k+1)\times 1}{\beta} + \underset{n\times 1}{\epsilon}$, $X = \underset{n\times p \quad k-p+1}{(X_1, X_2)}$, $\beta = \underset{p\times 1}{(\beta_1', \beta_2')'}$

① Including redundant variables does have a cost.

– the predictor using all variables will have a large variance than that based on X_1.

② On the other hand if $\beta_2 \neq 0$ & we use X_1, the prediction is biased.

• Suppose to predict $x'\beta = z = (x_1', x_2')'$ is $(k+1)$ vector

– Should we use $x_1'\tilde{\beta}_1$ based on X_1 or unbiased predictor $z'\hat{\beta}$ where $\hat{\beta}$ is the LSE of β based on X using all the exp. variable with $m=1$, the expected PE of the based predictor.

$$E(PE) = \sigma^2 + E\left[(x_1'\tilde{\beta}_1 - x'\beta)^2\right]$$

$$= \sigma^2 + \text{var}(x_1'\tilde{\beta}_1) + (x_1' E(\tilde{\beta}_1 - x'\beta)^2$$

$x' E(\tilde{\beta}_1) - x'\beta$ $\quad E(\tilde{\beta}_1) = \beta_1 + L\beta_2$

$= x_1'\beta_1 + x_1' L\beta_2 - x_1'\beta_1 - x_2'\beta_2 \quad L = (X_1'X_1)^{-1}X_1X_2$

$= (L'x_1 - x_2)'\beta_2$

$$E(PE) = \sigma^2 + \text{var}(x_1'\tilde{\beta}_1) + E\left[(L'x_1 - x_2)'\beta_2\right]^2$$

↳ for biased predictor.

For the unbiased predictor:

$$E(PE) = \sigma^2 + E\left(\underline{x}'\hat{\underline{\beta}} - \underline{x}'\underline{\beta}\right)^2$$

$$= \sigma^2 + \text{Var}\left(\underline{x}'\hat{\underline{\beta}}\right)$$

$$= \sigma^2 + \text{Var}\left(\underline{x}_1'\tilde{\underline{\beta}}_1\right) + \sigma^2\left(L'\underline{x}_1 - \underline{x}_2\right)' M \left(L'\underline{x}_1 - \underline{x}_2\right)$$

When $M = \left(X_2'(I_n - P_1)X_2\right)^{-1}$; $P_1 = X_1(X_1'X_1)^{-1}X_1'$

write $L'\underline{x}_1 - \underline{x}_2 = \underline{h}$

Comparing two PE's, the biased predictor $\underline{x}_1'\tilde{\underline{\beta}}_1$ has a smaller expected PE than the unbiased predictor if

$$\left(\underline{h}'\underline{\beta}_2\right)^2 < \sigma^2 \underline{h}'\mu\underline{h} \quad \text{——} (*)(*)$$

Since μ is p.d. using $(*)$

$$\left(\underline{h}'\underline{\beta}_2\right)^2 \le \left(\underline{h}'\mu\underline{h}\right)\underline{\beta}_2'\mu^{-1}\underline{\beta}_2$$

$\forall$ vectors $\underline{h}$

So if $\underline{\beta}_2'\mu^{-1}\underline{\beta}_2 = \underline{\beta}_2' X_2'(I_n - P)X_2\underline{\beta}_2$ (2)

$$< \sigma^2$$

then $(*)(*)$ is true for all $\underline{h}$.

$r'L$ is p.d. for any $\underline{b}$

$$\max_{u,\, h \neq 0} \frac{(\underline{h}'\underline{b})^2}{\underline{h}'L\underline{h}} = \underline{b}'L^{-1}\underline{b} \quad \text{——} (*)$$

— So the biased predictor will have the smaller expected PE.

Choosing the Best subset

1. Goodness of fit - criterion

RSS is a measure of goodness of fit, but it is not a good absolute measure, since we have two sets of variables S_1 and S_2, $S_1 \subset S_2$

Then the RSS for model based on S_2 will be smaller than for the model based on S_1. (provided the variables are not known combination of those in S_1)

- A way of correcting the RSS to accurate for the no of residuals is to use the estimated residual variance.

$$S^2 = \frac{RSS}{n-p}, \quad p \text{ is no of columns in } X.$$

- Another criterion is adjusted R^2

$$\bar{R}^2 = R_{adj}^2 = 1 - \frac{n-1}{n-p}\cdot(1-R^2)$$

- Consider the model having the largest $\bar{R}^2$

The use of $\bar{R}^2$ is motivated by testing the accuracy of the model as the 'full model' consider of all available exp variables.

- Suppose that there are K exp variables and would under consideration was $p-1 < K$ variable.

- R_p^2 & R_{K+1}^2 : coeff of determination
- F test for testing the accuracy of the small model as the larger model is

$$H_0: \beta_p = \beta_{p+1} = \dots = \beta_K = 0$$

The f-Statistic.

$$F_p = \frac{R_{K+1}^2 - R_p^2}{1-R_{K+1}^2} \cdot \frac{n-K-1}{K+1-p}$$

$$F = \frac{(RSS_H - RSS)/q}{RSS/(n-p)}$$

$$RSS = (1-R^2)\sum(y_i-\bar{y})^2$$

So, $1-R_p^2 = (1-R_{K+1}^2)\dfrac{(K+1-p)f_p + (n-K-p)}{n-K-1}$

and $\bar{R}_p^2 = 1 - (1 - R_p^2)\frac{n-1}{n-p}$

$$= 1 - \left(1 - R_{K+1}^2\right)\frac{(K+1-p)F_p + (n-K-1)}{n-K-1} \cdot \frac{n-1}{n-p}$$

If $F_p \geq 1$; $\frac{(K+1-p)F_p + (n-K-1)}{n-K-1} \geq \frac{K+1-p+n-K-1}{n-K-1}$

So, $\bar{R}_p^2 \leq 1 - \left(1 - R_{K+1}^2\right)\frac{n-1}{n-K-1}$

$$= \bar{R}_{K+1}^2$$

large values of F_p are evidence in favour of K variables model. This motivates us to use adjusted R^2 as an model selection criterion.

Note:

$$\bar{R}_p^2 = 1 - (1 - R_p^2)\frac{n-1}{n-p}$$

$$; S_p^2 = \frac{RSS_p}{n-p}$$

$$\bar{R}_p^2 = 1 - \frac{(n-1)S_p^2}{\sum(y_i - \bar{y})^2}$$

So, the model is max^m $\bar{R}^2$ is also the model with minimum S^2.

Model Selection Criterion based on Prediction Error

~~To get an operational criterion~~

$$ME = \| \mu_0 - X_0 \hat{\beta} \|^2$$

To get an operational criterion we must ~~expected~~ estimate the expected ME.

Mallow's @ C_p

RSS_p: RSS from fitting a model with p-parameters be $n \times p$ regression matrix X_p.

If $P_p = X_p (X_p' X_p)^{-1} X_p'$, then

$$E[RSS_p] = E[y'(I - P_p)'y)] = \mu'(I-P)\mu + (n-p)\sigma^2$$

$= E(ME) + (n-2p)\sigma^2$; $E(ME) = \underline{\mu}'(I-P)\underline{\mu} + \sigma^2 p$.

$$\frac{E(ME)}{\sigma^2} = \frac{E(RSS_p)}{\sigma^2} + 2p - n$$

If we had an estimate $\hat{\sigma}^2$, then we could use it & $C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n$ → as an estimate of $E(ME)/\sigma^2$, the scaled expected model error.

- If the model fits model in the sense that $\underline{\mu}$ is well-approximated by vectors in $\mathcal{C}(X_p)$, then $\underline{\mu}'(I_n - P_p)\underline{\mu} = \|(I_n - P_p)\underline{\mu}\|^2$ will be small.

$$E(C_p) \approx \frac{E[RSS_p]}{\sigma^2} + 2p - n.$$

$$= \frac{\underline{\mu}'(I_n - P_p)\underline{\mu}}{\sigma^2} + \frac{(n-p)\sigma^2}{\sigma^2} + 2p - n$$

$$\approx p$$

So, the expected C_p will be close to p.

Mallow's suggested using C_p plot, a plot of C_p v/s p for all possible models.

→ consider model for which $C_p \approx p$ (i.e. the model fits well)

- It is a common practise to use all K variables to estimate σ^2.

$$\hat{\sigma}^2 = \frac{RSS_{K+1}}{n-K-1}$$

$$\Rightarrow \quad C_{k+1} = (k+1)\,\frac{RSS_p}{RSS_{k+1}} \cdot \frac{n-k-1}{n-p} = \frac{1-\bar{R}_p^2}{1-\bar{R}_{k+1}^2}$$

If n is large compare to p, the smallest value of $C_p - p$ corresponds approximately to the largest $\bar{R}_p^2$. (Show this)

Cross-Validation (CV)

— (x_{oi}, y_{oi}) ; $i = 1, 2, .., m$ → a new data following the same model as the training set data (x_i, y_i), $i = 1, 2, .., n$

— An obvious measure of PE

$$\frac{1}{n} \sum_{i=1}^{} \left[y_{oi} - \underline{x}_{oi}' \hat{\underline{\beta}} \right]^2$$

$\hat{\underline{\beta}}$: estimated using training set data.

- In practice, we donot have additional data
- Sometimes we can divide that two disjoint set of obs → one for training set & the other for prediction.
- What we do →
  - Select a subset D of d observations.
  - estimate the prediction using (n-d) obs^{n}s.
  - Then calculate ⊛ ;, [PE(D)]

Repeat this process for selected d subsets $D_1, D_2, \ldots$ and average the resulting estimate $PE(D_i)$.

↳ This process is called cross-validation.

The most popular choice of d is 1. (This involves leaving out one obs in turn and calculating the estimate

from the remaining $(n-1)$ obs).

The error in predicting the ith obs as $y_i - \underline{x}_i' \hat{\beta}(i)$.

↳ heads: the leave one-out or CV(1) prediction error estimating

$$CV(1) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \underline{x}_i' \hat{\underline{\beta}}(i)\right)^2$$

$$y_i - \underline{x}_i' \hat{\underline{\beta}}(i) = y_i - \underline{x}_i' \left[\hat{\underline{\beta}} - \frac{(X'X)^{-1} \underline{x}_i (y_i - \underline{x}_i' \hat{\underline{\beta}})}{1-h_i}\right]$$

$$= y_i - x_i' \hat{\underline{\beta}} + \frac{h_i (y_i - x_i' \hat{\underline{\beta}})}{1-h_i}$$

$$= \frac{y_i - x_i' \hat{\beta}}{1-h_i}$$

$$CV(1) = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - x_i' \hat{\beta})^2}{(1-h_i)^2}$$

Expected value of CV(1):

write $\{(I_n - P)\underline{\mu}\}_i = \eta_i$ & $\underline{\epsilon} = \underline{y} - \underline{\mu}$

$$y_i - \underline{x}_i' \hat{\beta} = \{\underline{y} - X\hat{\beta}\}_i$$

$$= \{(I_n - P)\underline{y}\}_i$$

$$= \{(I_n - P)\underline{\mu}\}_i + \{(I - P)\underline{\epsilon}\}_i$$

$$= \eta_i + \{(I_n - P)\underline{\epsilon}\}_i$$

$$E\left[(y_i - \underline{x}_i' \hat{\underline{\beta}})^2\right] = E\left[\left(\eta_i + \{(I-P)\underline{\epsilon}\}_i\right)^2\right]$$

$$= \eta_i^2 + E\left[\{(I-P)\underline{\epsilon}\}_i^2\right]$$

Let, D_i : be a diagonal matrix whoose ith diagonal element is 1 and rest are zero.

$$\left\{(I_n - P)\underline{\epsilon}\right\}_i^2 = \underline{\epsilon}'(I_n - P)' D_i (I_n - P)\underline{\epsilon}$$

$$E\left[\left\{(I_n - P)\underline{\epsilon}\right\}_i^2\right] = E\left[\underline{\epsilon}'(I_n - P) D_i (I_n - P)\underline{\epsilon}\right]$$

$$= 0 + \sigma^2 \, tr\left((I_n - P) D_i (I_n - P)\right]$$

$$= \sigma^2 tr\left(D_i (I_n - P)\right)$$

$$= \sigma^2 (1 - h_i)$$

$$E\left[(y_i - x_i'\hat{\beta})^2\right] = \eta_i^2 + \sigma^2(1 - h_i)$$

$$E[nCV(1)] = \sum_{i=1}^{n}\left[\frac{\eta_i^2 + \sigma^2(1-h_i)}{(1-h_i)^2}\right]$$

The expected PE

$$E(PE) = (n+p)\sigma^2 + \underline{\mu}'(I-P)\underline{\mu}$$

$$= (n+p)\sigma^2 + \|(I_n - P)\underline{\mu}\|^2$$

$$= (n+p)\sigma^2 + \sum_{i=1}^{n}\eta_i^2$$

$$= \sum_{i=1}^{n}\left[\sigma^2(1+h_i) + \eta_i^2\right]$$

$$E[nCV(1) - PE] = \sum_{i=1}^{n}\left[\frac{\eta_i^2 + (1-h_i)\sigma^2}{(1-h_i)^2} - \sigma^2(1+h_i) + \eta_i^2\right]$$

$$= \sum_{i=1}^{n}\eta_i^2 \frac{h_i(2-h_i)}{(1-h_i)^2} + \sigma^2\sum_{i=1}^{n}\frac{h_i^2}{1-h_i}$$

≥ 0 $\qquad$ ≥ 0

Both the terms are positive & $nCV(1)$ tends to overestimate the PE.

$CV(d) =$

AIC → A Information Criterion.

BIC →

CV for $d > 1$

Suppose $D = \{i_1, \ldots, i_d\}$ are the cases to be omitted,

Let, $\underline{Y}_D' = \{Y_{i_1}, \ldots Y_{i_d}\}$

$\hat{\underline{\beta}}_{(D)}$: estimate of $\underline{\beta}$ based on the obs^n not in D.

and $H_D = X_D (X'X)^{-1} X_D'$

Since we are predicting the cases in D simultaneously and comparing them with $\underline{Y}_D$:

$$\underline{Y}_D - X_D \hat{\underline{\beta}}_{(D)} = (I - H_D)^{-1} (\underline{Y}_D - X_D \hat{\underline{\beta}})$$

So, $$CV(d) = \binom{n}{d}^{-1} \sum_D (Y_D - X_D \hat{\beta})' (I - H_D)^{-2} (\underline{Y}_D - X_D \hat{\underline{\beta}})$$

where Sum is taken over all d subsets of observation.

- For even $d = 2$ or 3, it is computationally very inclusive

Estimating Distributional Discrepancy: (AIC)

This criterion is based on a discrepancy between the true dist^n of the data $\underline{Y}$ and the dist^n specified by the model under consideration.

- $f(\underline{y})$: density of the true dist^n.
- $g(\underline{y})$: " " specified model.

Discrepancy measure in the Kullback-Leibler discrepancy.

$$KL(f,g) = \int \log \frac{f(\underline{y})}{g(\underline{y})} f(\underline{y})\, d\underline{y} = \int \log(f(\underline{y}))\, f(\underline{y})\, d\underline{y} - \int \log g(\underline{y})\, f(\underline{y})\, d\underline{y}$$

Note: This is not a true distance measure in the sense

$$KL(f,g) \neq KL(g,f)$$

$$KL(f,g) \geq KL(f,f) = 0$$

In practise, we are interested in a family of models $g(\underline{y};\underline{\theta})$ where $\underline{\theta}$ ranges over some parameter space

→ use $KL(f,g;\underline{\theta})$

- The true model may or may not be of the form of $g(\underline{y};\underline{\theta})$ for some particular $\underline{\theta}$.
- We want to choose the model corresponding to the value of $\underline{\theta}$, that minimizes $KL(f,g;\theta)$ or equivalently, that minimizes

$$-\int \log g(\underline{y};\theta) f(\underline{y})\, d\underline{y} = -E\left[\log\left(g(\underline{y};\theta)\right)\right] \quad \text{——— (1)}$$

This is equal to $KL(f,g,\theta)$ upto a constant

$$\int \log f(\underline{y})\, f(\underline{y})\, d\underline{y} \rightarrow \text{indpt of } g.$$

(1) depends on two unknowns, the parameter θ & the true dist^n. To estimate θ

Suppose we have a sample Y and we estimate $\underline{\theta}$ using ML estimate $\hat{\underline{\theta}}$

- Let, $\hat{\theta}(y)$ be the MLE that maximizes $g(y,\theta)$ as a function of $\underline{\theta}$.

→ leads to the modified criterion.

$$\Delta = -\int \log\left(g(z);\ \hat{\theta}(y)\right) f(z)\,dz$$

Then $E(\Delta) = -\iint \log\left(g(z);\ \hat{\theta}(y)\right) f(z)\, f(y)\, dz\, dy$

↳ Still f is unknown.

To get operational criterion we need to estimate Δ.

- The standard estimate of Δ [basically 2Δ] is the Alkaike Information Criterion (AIC)

$$AIC = -2\log\left(g(y;\hat{\theta}(y)\right) + 2r$$

Where r is the dimension of the par. vector θ.

Incase of linear models,

- the true models $f(y)$ of y may be taken as multivariate normal with mean vector μ and variance matrix $\sigma_0^2 I_n$.

A typical considered model $g(y,\theta)$ is also multivariate normal with mean vector $X\beta$ and variance matrix $\sigma^2 I_n$.

By considerate model g,

$$\log g(y,\theta) = \frac{n}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)$$

& the MLEs are; $\tilde{\beta} = (X'X)^{-1}X'Y$

$$\tilde{\sigma}^2 = \frac{RSS}{n}$$

Let Y_0 have the same distn as Y and is indpt of Y

Then $\Delta = E_{Y_0}\left[-\ln g(Y_0;\hat{\theta}(Y))\right]$

$$= \frac{n}{2}\ln(2\pi\hat{\sigma}^2) + \frac{1}{2\hat{\sigma}^2} E_{Y_0}\left[(Y_0 - X\hat{\beta})'(Y_0 - X\hat{\beta})\right]$$

$$= \frac{n}{2} \ln\left(2\pi\hat{\sigma}^2\right) + \frac{1}{2\hat{\sigma}^2}\left[n\sigma_0^2 + \|\underline{\mu} - X\hat{\underline{\beta}}\|^2\right]$$

Note that, MLEs $\tilde{\beta}$ and $\tilde{\sigma}^2$ are indp $\underline{y}$. and

write $\lambda = \dfrac{\underline{\mu}'(I-P)\underline{\mu}}{\sigma_0^2}$

$$E_Y[2\Delta] = E_Y\left[n\log\left(2\pi\hat{\sigma}^2\right) + \frac{1}{\hat{\sigma}^2}\right]\left\{n\sigma_0^2 + \|\underline{\mu} - X\hat{\underline{\beta}}\|^2\right\}$$

$$= nE\left[\log 2\pi\hat{\sigma}^2\right] + \left\{n\sigma_0^2 + E\left(\|\underline{\mu} - X\hat{\underline{\beta}}\|^2\right)\right\}E\left[\frac{1}{\hat{\sigma}^2}\right]$$

$$= nE\left(\log 2\pi\hat{\sigma}^2\right) + \left\{n\sigma_0^2 + \sigma_0^2 p + \sigma_0^2\lambda\right\}E\left(\frac{1}{\hat{\sigma}^2}\right)$$

$$= nE\left[\log 2\pi\hat{\sigma}^2\right] + (n+p+\lambda)\,E\left[\frac{\sigma_0^2}{\hat{\sigma}^2}\right] \quad \text{---------- (2)}$$

To estimate this we use AIC.

Take $r = p+1$

$$AIC = n\log\left(2\pi\hat{\sigma}^2\right) + \frac{\left(\underline{y} - X\hat{\underline{\beta}}\right)'\left(\underline{y} - X\hat{\underline{\beta}}\right)}{\hat{\sigma}^2} + 2(p+1)$$

$$= n\log\left(2\pi\hat{\sigma}^2\right) + n + 2(p+1)$$

$$= n\log\left(2\pi\hat{\sigma}^2\right) + n + 2(p+1) \quad \text{---------- (3)}$$

$$(n+p+\lambda)\left(1+O(n^{-1})\right) = n+p+\lambda + \frac{n+p+\lambda}{n}O(1) = n+p+\lambda+O(1)$$

Comparing two expression (2) & (3) & using the approximation $E\left[\sigma_0^2/\hat{\sigma}^2\right] = 1 + O\left(n^{-1}\right)$

So upto $O(1)$, explicit value of AIC and $E[2\Delta]$ are the same.

↳ suggests that AIC is a reasonable of $E[2\Delta]$

$$\left[x_n = O\left(\frac{1}{n}\right) \Rightarrow n x_n = O(1)\right]$$