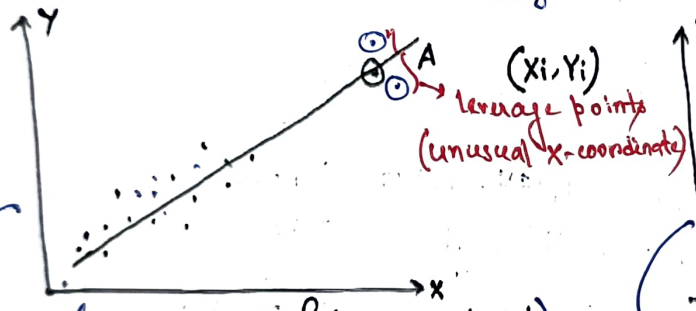
CHAPTER 6: MODEL ADEQUACY CHECKING

Residual: $e_i = y_i - \hat{y}_i$ — y_i is the observation, $\hat{y}_i$ is the corresponding fitted value.

Check the assumption: $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

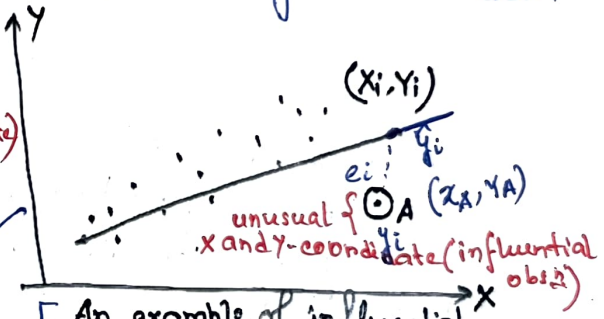
- ϵ_i's are independent but residuals e_i's are not independent as the n residuals have only $(n-k)$ DF. (all the residuals cannot be chosen independently in MLR)
- It is convenient to think of the residuals as the observed value of the errors. That's why we test the underlying assumptions on residuals.
- Plotting the residuals is an effective way to investigate how well the reg. model fit the data or to check the model assumption. [several residuals plots to be made to verify it]

Leverage & influential obsn.s: **Leverage pt:** A pt. that has unusual x-coordinate from the rest of the observations.
Influential pt: A pt. that has moderately unusual x-coordinate & y-value is unusual as well.



[An example of leverage point. The unusual pt. A is on the trend of the data set.]

[$L_i > \frac{2p}{n}$ suggests leverage point (software Rule)
p: No. of parameters in your model
n: total no. of observation.]



[An example of influential observation that has noticeable impact on the model coefficients.]

[For influential, $D_i > 1$ (Cook's statistic) shows influential obsn.]

Various Types of Residuals

(a) Regular residuals e_i
(b) Standardize residuals d_i
(c) Studentized residuals r_i
(d) PRESS residuals $e_{(i)}$

- Outlier Check: using Box plot (Descriptive statistics)

Residual plots

1. Normal Probability plot (Checking normality assumption)
2. Plot of residuals (e_i) against the fitted values ($\hat{y}_i$) (Checking constant variance assumption)
3. Partial regression & partial residual plot.

RATIONALE — **Residual Analysis**

SLR: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$; $i = 1(1)n$
MLR: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{k-1} x_{ik-1} + \epsilon_i$

— Here we present several techniques to check these assumptions on errors. (stability & adequacy checking)

Assumptions:

$E(\epsilon_i) = 0$
$V(\epsilon_i) = \sigma^2$
$\epsilon_i \overset{IID}{\sim} N(0, \sigma^2)$
[errors are uncorrelated]

The hat matrix & the various types of residuals:

MLR : $Y = X\beta + \epsilon$; $V(\epsilon) = \sigma^2 I_n$.

Solution: $\hat{\beta} = (X'X)^{-1}X'Y$ if $(X'X)$ is non-singular.

Fitted model $\hat{Y} = X\hat{\beta}$

$= X(X'X)^{-1}X'Y$

$= HY$, say, where $H = X(X'X)^{-1}X'$

= Hat matrix

(It maps Y to $\hat{Y}$, so called Hat matrix)

Hat Matrix: $H = ((h_{ij})) = \begin{pmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ \dots & \dots & \dots & \dots \\ h_{n1} & h_{n2} & \dots & h_{nn} \end{pmatrix}$

PROPERTIES OF HAT MATRIX:

H is symmetric, i.e. $H = H^T$

H is idempotent, i.e. $H^2 = H$

} easy to check.

(a) Regular Residual:-

$e = Y - \hat{Y} = Y - HY = (I-H)Y \quad \because \hat{Y} = HY$

$= (I-H)(X\beta + \epsilon)$

$= X\beta - HX\beta + (I-H)\epsilon$

$= X\beta - X(X'X)^{-1}X'X\beta + (I-H)\epsilon = X\beta - X\beta + (I-H)\epsilon$

$\therefore \boxed{e = (I-H)\epsilon}$

Variance - covariance matrix of e: $Var(e) = (I-H)\sigma^2 I (I-H)$ $[V(\epsilon) = \sigma^2 I]$

$= \sigma^2 (I-H)^2$ [since $H^2 = H \Rightarrow (I-H)^2 = (I-H)$]

$= \sigma^2 (I-H)$

$e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$ $\therefore V(e_i) = \sigma^2(1 - h_{ii})$; where h_{ii} is the i^{th} diagonal element of the hat matrix H.

Thus, $Cov(e_i, e_j) = -\sigma^2 h_{ij}$.

Hat matrix: $H = X(X'X)^{-1}X'$ $\quad X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix}$

Diagonal element: $h_{ii} = x_i'(X'X)^{-1}x_i$;

where x_i' is the i^{th} row of X matrix. (Recall from MLR)

Note that h_{ii} measures the distance of i^{th} observation from the center of x-coordinate and $0 \le h_{ii} \le 1$. Important note: h_{ii} is large if the i^{th} observation is a leverage point, so, from h_{ii}, we can get information on leverage pts. in the data.

(b) Studentized residuals:

We define $r_i = \dfrac{e_i}{\sqrt{MS_{res}(1-h_{ii})}}$ $\quad V(e_i) = \sigma^2(1-h_{ii})$ $\quad \hat{\sigma}^2 = MS_{Res}$

Studentized residuals have constant variance, i.e., $V(r_i) = 1$ regardless of the location in x-coordinate when the form of the model is correct.

[For outliers in the data, (ith influential/leverage observation) h_{ii} is large, i.e. close to 1, so $r_i >> d_i$.]

(c) Standardized residuals:

Define $d_i = \dfrac{e_i}{\sqrt{MS_{res}}}$; we approx. MS_{Res} as variance of ith residual e_i.

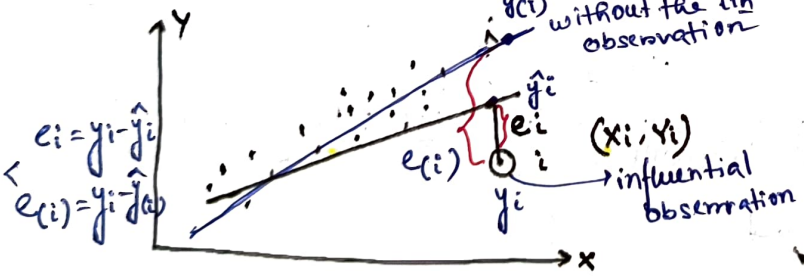
For influential observation h_{ii} is large ; $0 \le h_{ii} \le 1$, case of studentized residuals. For standardized residuals, $h_{ii} = 0$.

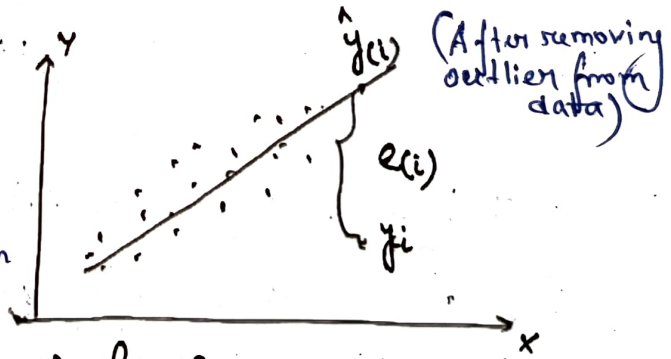
- Example of Studentized and Standardized residual: Delivery time data in Montgomery, Peck, Vining.

(d) PRESS Residual

ith press residual $e_{(i)} = y_i - \hat{y}_{(i)}$

where, $\hat{y}_{(i)}$ is the fitted value of ith response based on all observations except ith one.





- We delete ith observation (influential), fit the regression model to the remaining (n-1) observations, and predict y_i.
- It is possible to calculate PRESS residuals from the result of one single fit to all n obsn..

$$e_{(i)} = \frac{e_i}{1-h_{ii}} = y_i - \hat{y}_{(i)}.$$

[If ith observn is outlier, then $e_{(i)} - e_i \sim$ large.]

- Large PRESS residuals are useful in identifying obsn. where the model doesnot fit the data well.
- The PRESS statistic is $PRESS = \sum_1^n e_{(i)}^2 = \sum_1^n (y_i - \hat{y}_{(i)})^2$

$$= \sum \left(\frac{e_i}{1-h_{ii}}\right)^2.$$

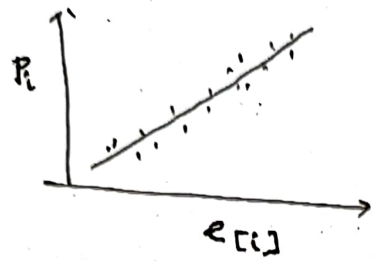
[PRESS statistic measures how well a reg. model will perform in predicting new data]

NOTE: Standardized residual and studentized residual almost give the same/similar information and have values very close to each other. The values are very different for any given information which is outlier (leverage/influential obsn).

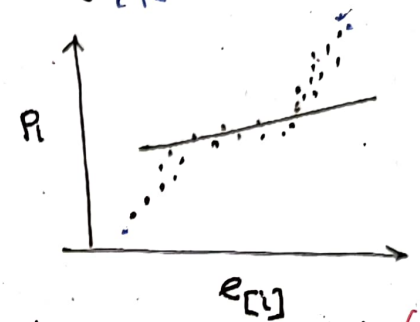
Residual Plots :-

Normal Probability Plot :- Let $e_1, e_2, \ldots, e_n$ be n residuals. Let $e_{[1]} < e_{[2]} < \cdots < e_{[n]}$ be the residuals ranked in increasing order.

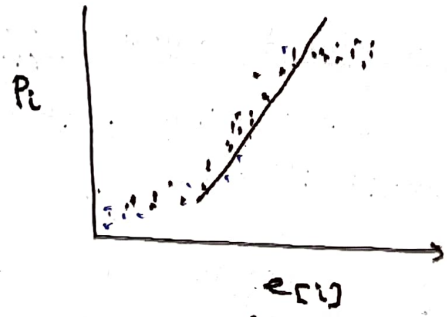
Plot $e_{[i]}$ Vs. cumulative probability $P_i = \frac{i - 1/2}{n}$ $\forall\, i = 1(1)n$ i is



(a) Normal distn. (Ideal situation)



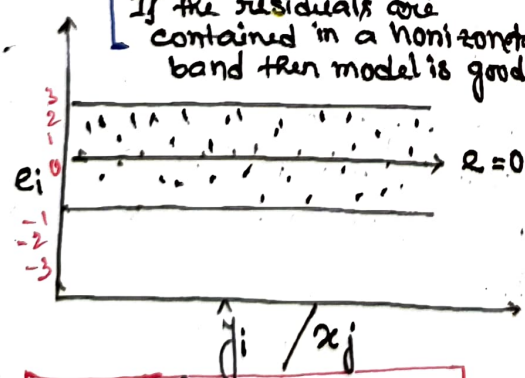
(b) heavy tailed distn. (Non-normal)



(c) light-tailed distn. (non-normal)

Plot of Residuals (e_i) Vs. fitted values ($\hat{y}_i$) / Regressor (x_j)

[If the residuals are contained in a horizontal band then model is good.] : TO CHECK

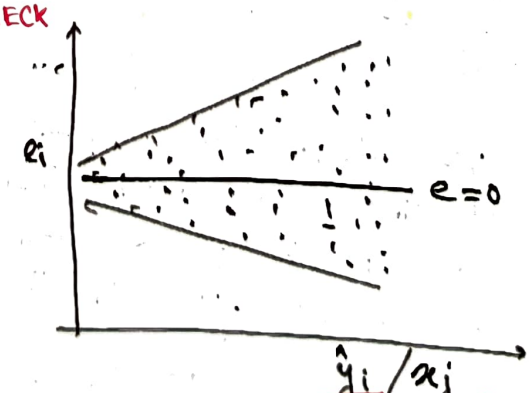


(a) Satisfactory model

"Good" reg. model will produce a scatter in residuals that is roughly constant with $\hat{y}$ and centered about $e=0$

$[V(\epsilon) = \sigma^2]$

i.e., constant variance



(b) Unsatisfactory model

Outward-opening Funnel

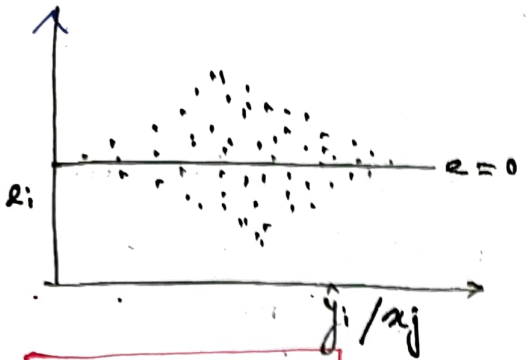
$V(\epsilon) \uparrow$ as $\hat{y} \uparrow$

For inward-open funnel

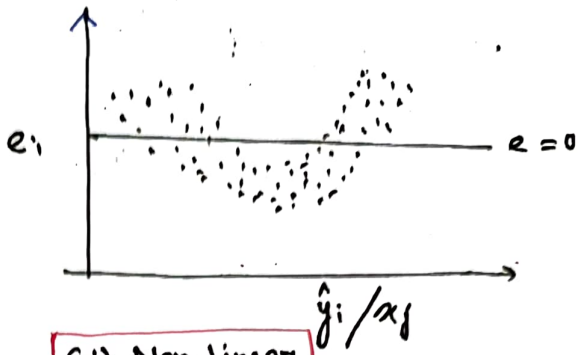
$V(\epsilon) \downarrow$ as $\hat{y} \uparrow$

$[V(\epsilon) \neq \sigma^2]$

i.e., non-constant variance



(c) Double-bow indicates non-constant variance, $V(\epsilon) \neq \sigma^2$
y is a proportion; $0 \le y \le 1$



(d) Non-linear
Other regresson variables are needed in the model. Consider extra term (square term x^2) to the model or transform Y.

→ Why do we plot the residuals e_i against the $\hat{y}_i$ and not against y_i for the usual linear model?

(Interested to find the relationship between y_i and e_i)

Ans. e_i and y_i are usually correlated. whereas e_i and $\hat{y}_i$ are not correlated. We assume a linear relationship between y_i and e_i. Suppose

SLR:
$$e_i = \beta_0 + \beta_1 y_i + \epsilon_i$$

→ [since $\sum \bar{e}(y_i - \bar{y}) = 0$]

LSE:
$$\hat{\beta}_1 = \frac{S_{ey}}{S_{yy}} = \frac{\sum (e_i - \bar{e})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2} = \frac{\sum e_i (y_i - \bar{y})}{SS_T} = \frac{\sum e_i y_i}{SS_T} \quad [\text{since } \sum e_i \bar{y} = 0]$$

$$= \frac{Y'e}{SS_T} = \frac{Y'(I-H)Y}{SS_T}$$

$$= \frac{Y'(I-H)(I-H)Y}{SS_T} \quad ; \ (I-H) \text{ is idempotent mtx.}$$

$$= \frac{e'e}{SS_T} = \frac{SS_{Res}}{SS_T}$$

$$= 1 - \frac{SS_{Reg}}{SS_T}$$

$$= 1 - R^2 \quad (R^2: \text{coefficient of multiple determination})$$

- There is a linear relationship between y_i and e_i and slope is $(1-R^2)$.
- We now check whether there is any linear relationship between e_i and $\hat{y}_i$. We can prove that the value of the slope $= 0$ in this case.

SLR: $e_i = \beta_0 + \beta_1 \hat{y}_i + e$ (assuming)

LSE $\hat{\beta} = \frac{S_{e\hat{Y}}}{S_{\hat{Y}\hat{Y}}}$

$$S_{e\hat{Y}} = \sum (e_i - \bar{e})(\hat{Y}_i - \hat{Y}) = \sum e_i \hat{Y}_i = e'\hat{Y} = Y'(I-H)HY$$

$$= Y'(H - H^2)Y \quad [\because H^2 = H] \quad \left[\begin{matrix} e = (I-H)Y \\ \hat{Y} = HY \end{matrix}\right]$$

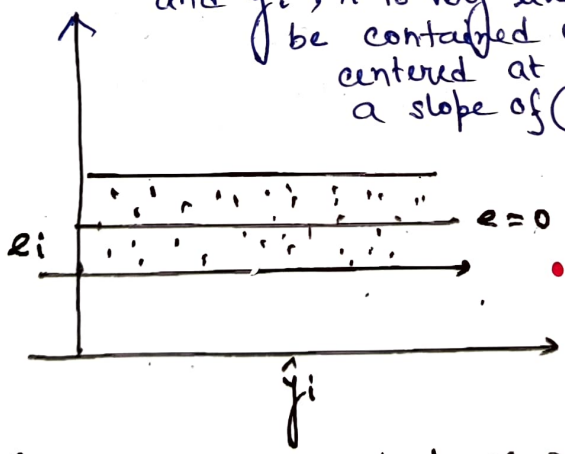
$$= Y'0Y \text{ , since H is idempotent.}$$

$$= 0$$

So, e_i is not linearly related with $\hat{y}_i$

- In case of e_i and Y_i, $\hat{\beta}_1 = 1 - R^2$.
- Unless $R^2 = 1$, there will be a slope of $(1-R^2)$ [positive slope]

in e_i Vs y_i plot, even if there is nothing wrong with the model, There is a theoretical relationship between residual and y_i, it is very likely that residuals will not be contained within a horizontal band centered at $e=0$. there will always be a slope of $(1-R^2)$ when $R^2 \neq 0$.



⇨ corresponding model fitted well.

- This is the justification of plotting $\hat{y}_i$ and e_i (there is no linear relationship between them).

- In case of plotting the residuals vs regressors, it may not show the marginal effect of a regressor x_j, given the other regressors in the model. (limitation of plotting e_i Vs x_j) [MLR case]

→ So, next we will discuss Partial Residual Plot.

[Show: $\hat{\beta}_1 = 1 - R^2$. [Alternative way] [Exercise 7.(b) of Tutorial Sheet]

→ MODEL: $e = a + bY$ $\quad Y = \beta_0 + \beta_1 x$

$$\beta_1 = \frac{S_{xy}}{S_{xx}}$$

$$b = \frac{e'Y}{Y'Y} = \frac{e'e}{Y'Y}$$

$$= \frac{SS_{Res}}{SS_T} = \frac{SS_T - SS_{Reg}}{SS_T} = 1 - R^2.]$$

NOTE: Plot of residual Vs. x_j (regressor) is important in determining relationship between the response variable Y and regressor x_j. In case of SLR, there is no difference between e_i Vs. $\hat{y}_i$ and e_i Vs. x. [SLR: $Y = \hat{\beta}_0 + \hat{\beta}_1 x$]

Partial Residual Plot :-

- Partial residual plot consider the marginal role of the reg. x_j given other reg. that are already in the model. [MLR case]
- In this plot, the response variable and the reg. x_j (say) are both regressed ag. the other regressors in the model (except x_j) residuals are obtained for each regression.
- The plot of these residuals against each other show the marginal role of reg. x_j on response variable y in the presence of other regressors in the model. [$y = f(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots)$ and $x_j = f(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots)$]

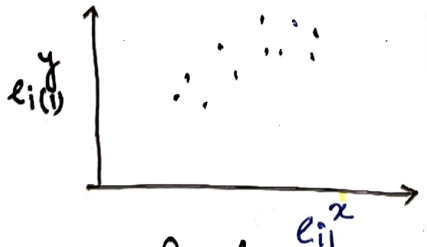
EXAMPLE: Consider the MLR model with two reg. X_1 and X_2. (Illustration)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

We are interested in marginal role of X_1 on response variable Y in the presence of X_2 in the model.

→ Regress Y on X_2 (except X_1): $\hat{y}_{i(1)} = \hat{\theta}_0 + \hat{\theta}_1 x_{i2}$; $e^{y}_{i(1)} = y_i - \hat{y}_{i(1)}$. [Eliminating effect of X_2 from Y to see the role of X_1]

→ Regress X_1 on X_2 : $\hat{x}_{i1} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{i2}$; $e^{x}_{i1} = x_{i1} - \hat{x}_{i1}$.



Is there any relationship between $e^{y}_{i(1)}$ and e^{x}_{i1} and then what pattern we expect from $e^{y}_{i(1)}$ vs. e^{x}_{i1} plot?

GENERAL CASE :-

The partial residual of y for x_j is defined as

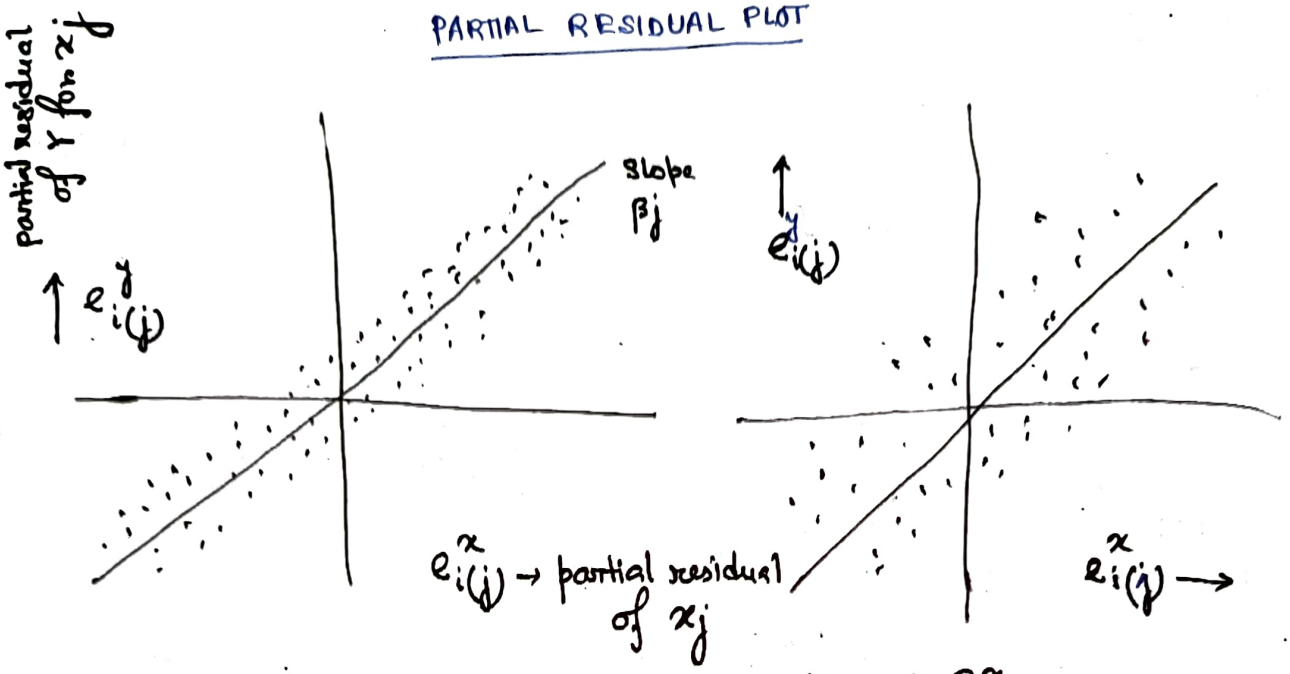
$$e^{y}_{i(j)} = y_i - \hat{y}_{i(j)}$$ for a regression model (MLR) with $(k-1)$ regressors,

where $\hat{y}_{i(j)}$ is a prediction of y_i from a reg. model using all regressors except x_j.

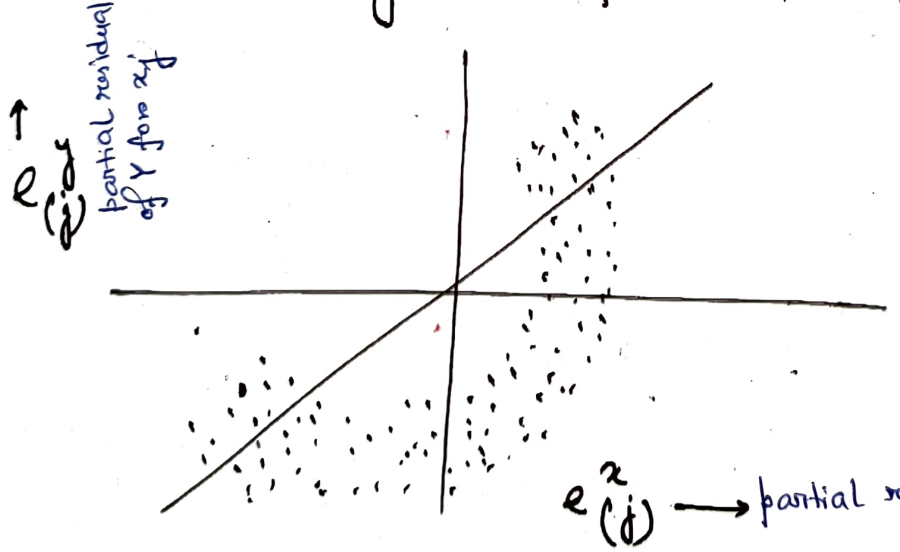
And $e^{y}_{i(j)}$ represents the variability in y_i not explained by a model that excludes the regressor x_j. (all regressors) How much variability can be explained by x_j alone?

- The partial residual of x_j is defined as $e^{x}_{i(j)} = x_{ij} - \hat{x}_{i(j)}$ [after regressing x_j of the remaining regressors]
  where, $\hat{x}_{i(j)}$ is a prediction of the reg. value x_{ij} from regression of x_j on all other reg. variables.
- $e^{x}_{i(j)}$ represents the variation in x_j that can NOT be explained by other regressors.
- $e^{y}_{i(j)} = \beta_j \cdot e^{x}_{i(j)} + \epsilon^{*}_{i}$: Partial residual plot should have a slope β_j (PROOF!) [If j^{th} regression is linearly related with Y then two residuals are also linearly related.

PARTIAL RESIDUAL PLOT



Partial residuals scatter around a line $y = \beta x$. x_j & Y.
Less scatter indicates strong relationship between



Curvilinear band: x_j is not linearly related to Y. Either we have to see for higher order term for x_j or transformation such as $\left(\frac{1}{x_j}, \log x_j\right)$ may be helpful. This is how we find the marginal role of regression variable on response Y.

Should influential obsn. be discarded?

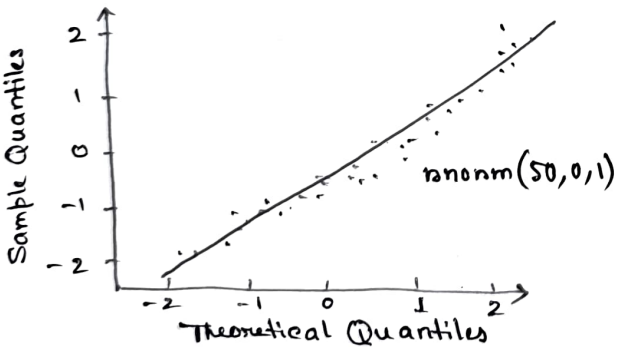
If there is an error in recording the obsn., then it can be discarded.

Checking Normality Assumption: We assume that all the errors are IID normal variables.

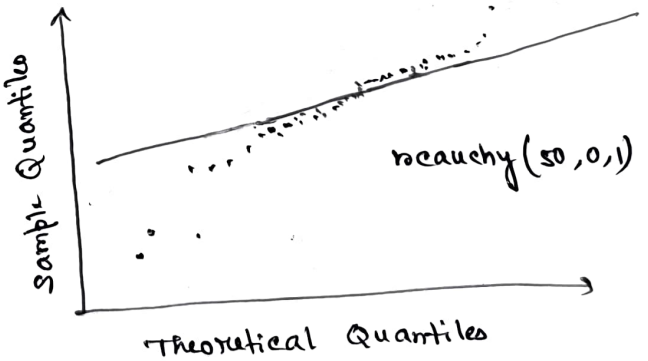
Graphical Approaches: (i) Normal Probability Plot; (already discussed)
(ii) Q-Q Plot;

Statistical Tests: (i) Kolmogorov-Smirnov Test
(ii) Anderson-Darling Test
(iii) Shapiro-Wilk Test

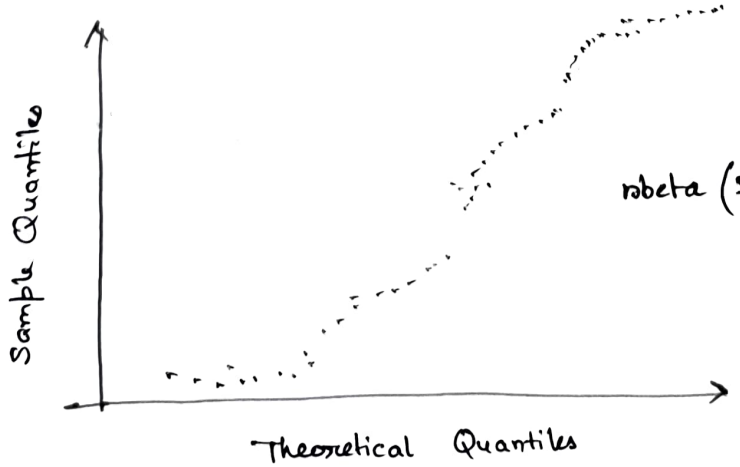
(ii) **Q-Q plot:** Q-Q plot is a graphical tool that is used to assess normality. It plots the sample quantiles (vertical axis) against the theoretical quantiles (horizontal axis) where the original data point x_i value is called sample quantile and the expected Z-score for the data point x_i is called theoretical quantile. If the data comes from a normal distribution, then the theoretical and sample quantiles match, hence giving an approximately straight line Q-Q plot. Otherwise, the plot is not straight line. Some Normal Q-Q plots are given below:



(a) Normal Distribution



(b) Heavy-tailed distribution



(c) Bi-modal distribution

TESTS FOR NORMALITY:

- **Kolmogorov-Smirnov Test:** Let $X_1, X_2, \ldots, X_n$ are assumed to come from a continuous distribution P. We want to test the following hypothesis:

  H_0: The samples come from P
  Vs. H_1: They donot come from P

  If F is the CDF of X under H_0, then the empirical distribution function F_{obs} is

  $$F_{obs} = \frac{I(X < x)}{\text{Total observations}}$$

  Let F_{exp} be the CDF associated with the null hypothesis. Then, the Kolmogorov-Smirnov statistic is given by $D_n = \max\{|F_{exp}(x) - F_{obs}(x)|\}$.

  Procedure:
  - The first step is to order the data. Let the ordered data be $x_{(1)}, \ldots, x_{(n)}$ so that $F_{obs}(x_{(i)}) = \frac{i}{n}$.
  - Find $F_{exp}(x_{(i)})$ for each i. Now tabulate the values of $|F_{exp}(x) - F_{obs}(x)|$ for x-value. Maximum of all these values is given by D_n.
  - Critical value for $\alpha = 0.05$ is given by $D_{c,0.05} = \frac{1.36}{\sqrt{n}}$.

  Hence, we reject the null hypothesis if $D_n > D_{c,0.05}$.

- **Shapiro-Wilk Test:** The Shapiro-Wilk test, proposed in 1965, calculates a W statistic that tests whether a random sample, $X_1, X_2 \ldots, X_n$ comes from a normal distribution. The W statistic is given by:

  $$W = \frac{\sum_{i=1}^{n} (a_i X_{(i)})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2},$$

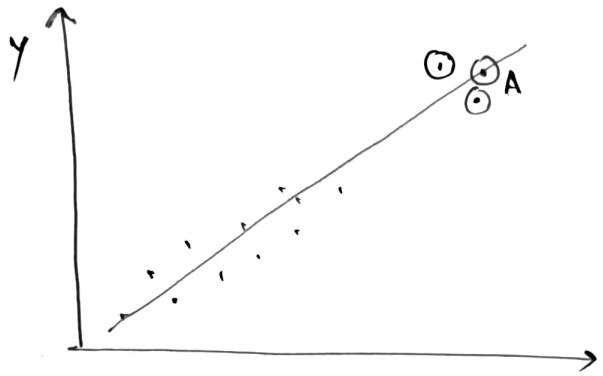
  where $X_{(i)}$ are the ordered sample values, $(a_1, \ldots, a_n) = \frac{m' V^{-1}}{c}$ for $c = \|V^{-1} m\|$. Here V is the variance-covariance matrix of the order statistics and $m = (E(X_{(1)}), \ldots, E(X_{(n)}))^T$.

- **Box-Cox Transformation:** If there is evidence of non-normality, then the standard remedy is to transform the response using Box-Cox transformation. The response variables in Box-Cox method need to be positive. We assume that there exists transformation parameter λ such that

  $$Y_I^{(i)} = g(Y_i, \lambda) = x_i'\beta + \epsilon, \text{ where } g(Y_i, \lambda) = \begin{cases} \frac{Y_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log Y_i & \text{if } \lambda = 0 \end{cases}$$

  Assuming that the transformed response $f(Y)$ follows multivariate normal distribution, we try to maximize the likelihood function of response variable w.r.t. λ. Details can be found in Montgomery book.

TEST FOR LEVERAGE POINT and INFLUENTIAL OBSERVATIONS: →



(X_i, Y_i) ; $i = 1(1)n$

– The point A has unusual x-coordinate from the rest of the observations.

Fig: An example of X leverage point

(X_i, Y_i) ; $i = 1(1)n$

– This point ⊙ has moderately unusual x-coordinate and the y-value is unusual as well.

Fig: An example of influential observation

Test for Leverage Point:

MLR model: $Y = X\beta + \epsilon$

LSE: $\hat{\beta} = (X'X)^{-1}X'Y$

Fitted model: $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY$ (say); where $H = X(X'X)^{-1}X'$

H is hat matrix that maps Y to $\hat{Y}$.

– Hat matrix plays an important roole in identifying leverage point.

– $H = X(X'X)^{-1}X'$

h_{ii} is the i^{th} diagonal element of the hat matrix H.

– $h_{ii} = x_i'(X'X)^{-1}x_i$ (standardized measure of the distance of the i^{th} observation from the center of x-coordinate)

$$X^{n\times k} = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix}$$

– We call a point leverage pt. if it has unusual x-coordinate.

– High h_{ii} value indicated i^{th} observation is a leverage point.

$$\therefore \bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{\text{tr}(H)}{n} = \frac{\text{rank}(H)}{n} = \frac{K}{n}.$$

As a general rule, $h_{ii} > 2\left(\frac{K}{n}\right)$ indicates that ith observation is a possible leverage point.

Test for Influential Observation:

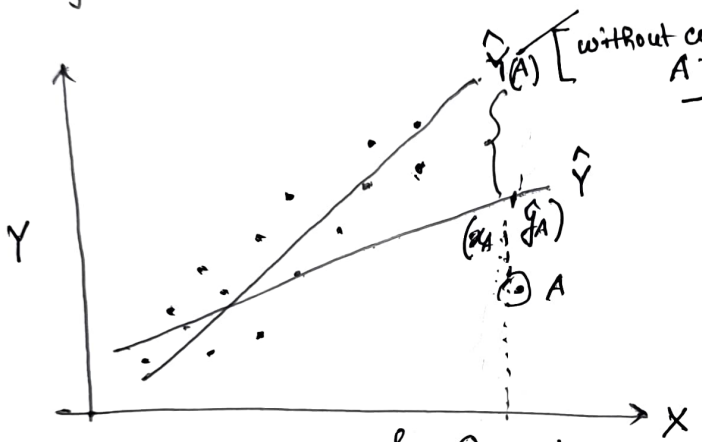


Fig: Example of Cook's statistic.

$$D_A = \underbrace{(\hat{Y} - \hat{Y}_{(A)})'(\hat{Y} - \hat{Y}_{(A)})}_{\text{euclidean distance}}$$

$$DFFITS_A = \hat{y}_A - \hat{y}_{(A)}$$

Definition of Outlier:

- An outlier is a data point whose response y does not follow the general trend of the rest of the data.
- A data point is said to be leverage if it has extreme predictor x values.
- A data point is said to be influential if it unduly influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results.
- It is important to note that the outliers and leverage datapoints have the potential to be influential, but we generally have to investigate further to determine whether or not they are actually influential.
- An influential point is one whose removal from the dataset would cause a large change in the fit. Three measures (popularly used) are as follows.

Determination of Influential Observation:-

Cook's statistic (D) for i^{th} obsn is based on the diff. between predicted response ($\hat{Y}$) obtained using all the obs. and predicted response $\hat{Y}_{(i)}$ obtained without the i^{th} obsn.

$$D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})' (\hat{Y}_{(i)} - \hat{Y})}{K MS_{Res}} ; \quad \hat{Y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix}, \quad \hat{Y}_{(i)} = \begin{pmatrix} \hat{y}_{(i)1} \\ \hat{y}_{(i)2} \\ \vdots \\ \hat{y}_{(i)n} \end{pmatrix}$$

$$= \frac{\sum (\hat{y}_{(i)j} - \hat{y}_j)^2}{K MS_{Res}}$$

- Squared Euclidean distance between the vector of fitted values and vector of fitted values when i^{th} obsn is deleted.

Compute: $D_1, D_2, \ldots, D_n$ (D_i : Cook distance for i^{th} obsn.)

- The value of D_i much larger than others indicates that i^{th} obsn. may be highly influential, preferably $D_i > 1$ is highly influential.

DFFITS (Difference between fit statistics) investigates deletion influence of the i^{th} observation on the fitted values. Similar to Cook's statistic but have different formula.

For the i^{th} obsn. this statistic is defined as

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{MS_{Res(i)} h_{ii}}} ;$$

where $\hat{y}_{(i)}$ is the fitted value of y_i obtained without the use of i^{th} obsn. $MS_{Res(i)}$ is the predicted value of MS_{Res} obtained without the use of i^{th} obsn.

- A possible high influential observation is indicated by

$$|DFFITS_i| > 2\left(\frac{K}{n}\right)^{1/2}$$

DFBETAS : How much reg. coeff. $\hat{\beta}_j$ changes, if the i^{th} obsn. is deleted. $DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{MS_{Res(i)} (X'X)^{-1}_{jj}}}$; $i = 1(1)n$, $j = 0(1)K-1$

(Difference between Betas)

$\hat{\beta}_{j(i)}$ is the j^{th} reg. coefficient computed without using i^{th} obsn. As a general rule.

A possible high influential obsn. is indicated by

$$|DFBETAS_{ij}| > \frac{2}{\sqrt{n}}.$$

Checking Errors Autocorrelation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i; \quad \text{Data: } (x_i, y_i);\ i = 1(1)n$$

Basic Assumptions: ① $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$, $cov(\epsilon_i, \epsilon_j) = 0$

② $\epsilon_i \overset{ind.}{\sim} N(0, \sigma^2)$

- When data (y_t, x_t) are collected sequentially in time the usual assumption of independence of errors is not guaranteed. Such data are called time series data.
- Errors are autocorrelated (serially correlated means correlation between errors and steps apart are always the same, i.e.,

$$Corr(\epsilon_t, \epsilon_{t+s}) = \rho_s \quad ; s = 1, 2, 3, \ldots$$

- Correlation between residuals one (or two or three) steps apart is called lag-1 (or 2 or 3) serial correlation.
- Source of Autocorrelation:

Primary source of autocorrelation in regression problem involving time series data is failure to include one or more important regressors(s) in the model.

- Effect of Autocorrelation:

1. If $Corr(\epsilon_i, \epsilon_j) \neq 0$,

then $Y = X\beta + \epsilon$

LSE: $\hat{\beta} = (X'X)^{-1}X'Y$ is unbiased but not having minimum variance.

i.e., If errors are correlated then $\hat{\beta}$ is not BLUE. $(V(\epsilon) = \sigma^2 V \neq \sigma^2 I)$

2. When the errors are positively autocorrelated then the MS_{Res} may seriously underestimate σ^2.

SLR: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (fitted model)

$V(\hat{\beta}_1) = \dfrac{MS_{Res}}{S_{xx}}$, standard error is $se(\hat{\beta}_1) = \sqrt{MS_{Res}/S_{xx}}$ (will be small). Confidence interval for β_1 will be small.

Example (Time Series Data)

SOFT DRINK CONCENTRATE DATA

YEAR	SALES (Y)	EXPENDITURES (X) (1K$)
1	3083	75
2	3149	78
3	3218	80
4	3239	82
5	3298	84
6	3374	88
7	3475	93
8	3569	104
9	3597	109
10	3725	115
⋮	⋮	⋮

SALES = f (EXPENDITURES)

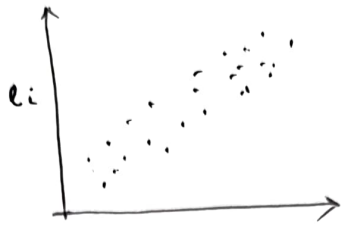
GROWTH OF POPULATION/POPULATION SIZE can be another variable need to be observed.

(Source of autocorrelation)

DETECTING AUTOCORRELATION:

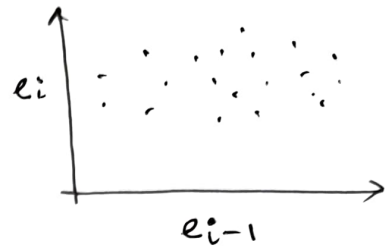
- Residual plot is useful for the detection of autocorrelation. Plot: (x_t, y_t)

[Residual of identical sign occur in a cluster then it indicates positive autocorrelation.]



(lag-1) $e_i = \hat{\rho} e_{i-1}$

[Lower left to upper right pattern indicates positive lag-1 autocorrelation. i.e., $\hat{\rho} > 0$.]



[errors are uncorrelated.]

SUMMARY ON MODEL ASSUMPTIONS:→

OLS (ordinary least square) is the 'BEST' procedure for estimating a linear regression model only under certain assumptions. Assumption that are required to hold are as follows:

[fixed X, stochastic Y]

A-1: The regression model is linear, correctly specified, and has an additive error term.
- linear in the coefficients.
- random error is additive
- having "right" independent variables.

$Y = X\beta + \epsilon$.

A-2: The error term has a zero population mean. $[E(\epsilon_i) = 0]$

A-3: All explanatory variables are uncorrelated with the error term. $[\text{cov}(x_i, \epsilon_i) = 0]$

A-4: Observations of the error term are uncorrelated with each other (no serial correlation). $[\text{cov}(\epsilon_i, \epsilon_j) = 0]$

A-5: The error term has a constant variance (no heteroscedasticity) $[\text{Var}(\epsilon_i) = \sigma^2]$

A-6: No explanatory variable is a perfect linear function of any other explanatory variables (no perfect multicollinearity). Also, if two or more variables are highly correlated, multicollinearity exist and could be problematic.

A-7: The error term is normally distributed. $[\epsilon_i \sim N(0, \sigma^2)]$

Durbin-Watson test:- (TEST FOR AUTOCORRELATION)

Suppose we wish to fit the model $y_u = \beta_0 + \sum \beta_i x_{iu} + \epsilon_u$, $u = 1(1)n$ by LS technique to obsn $(y_u, x_{1u}, x_{2u}, \ldots, x_{ku})$

We usually assume $\epsilon_u \overset{iid}{\sim} N(0, \sigma^2)$: $\rho_s = 0$ (s-step apart correlation between errors $= 0$)

We want to see if this assumption is justified / NOT.

Test: $H_0: \rho_1 = 0$ Vs. $H_1: \rho_s = \rho^s$ $(\rho \neq 0, |\rho| < 1)$

(H_0: No autocorrelation) comes from the assumption that

$$\epsilon_u = \rho \epsilon_{u-1} + z_u$$ [first order autoregressive errors]

where $z_u \sim N(0, \sigma^2)$ & is independent of $\epsilon_{u-1}, \epsilon_{u-2}, \ldots$ & of $z_{u-1}, z_{u-2}, \ldots$

$$\epsilon_u = \rho \epsilon_{u-1} + z_u$$

$$= \rho(\rho \epsilon_{u-2} + z_{u-1}) + z_u$$

$$= \rho^2 \epsilon_{u-2} + \rho z_{u-1} + z_u$$

$$= \rho^2(\rho \epsilon_{u-3} + z_{u-2}) + \rho z_{u-1} + z_u$$

$$= \rho^3 \epsilon_{u-3} + \rho^2 z_{u-2} + \rho z_{u-1} + z_u$$

$$= \sum_{k=0}^{u} \rho^k z_{u-k}$$

$$E(\epsilon_u) = 0 \; ; \quad V(\epsilon_u) = (1 + \rho^2 + \rho^4 + \ldots)\sigma^2$$

$$= \frac{\sigma^2}{1-\rho^2}$$

$$Cov(\epsilon_u, \epsilon_{s+u}) = \rho^{|s|} \cdot \sigma^2 \cdot \frac{1}{1-\rho^2}$$

$$Corr(\epsilon_u, \epsilon_{s+u}) = \rho^{|s|}.$$

$$\epsilon_u \sim N\left(0, \frac{\sigma^2}{1-\rho^2}\right)$$

Under H_0: $\rho = 0$, $\epsilon_u \overset{ind.}{\sim} N(0, \sigma^2)$.

Test: $H_0: \rho_s = 0$ Vs. $H_1: \rho_s = \rho^{|s|}$

— To test, we fit the model $Y = X\beta + \epsilon$ & compute the residuals e_i, then Durbin-Watson statistic is

$$d = \frac{\sum_{u=2}^{n} (e_u - e_{u-1})^2}{\sum_{u=1}^{n} e_u^2}$$

— The distn of d lies between 0 & 4 and symmetric about 2.

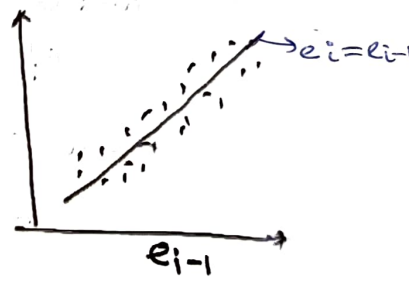
One sided test against the alternative:-

① $H_0: \rho = 0$ Vs $H_1: \rho > 0$

If $d < d_L$ reject H_0

If $d > d_U$ accept H_0

If $d_L < d < d_U$, test is inconclusive.



Thus, positive autocorrelation indicates successive error term are of similar magnitude & the diff. in residuals $e_i - e_{i-1}$ will be small. Then d is small. Thus, we reject H_0.

② $H_0: \rho = 0$ Vs. $H_1: \rho < 0$

If $4 - d < d_L$ reject H_0

If $4 - d > d_U$ accept H_0

If $d_L < 4 - d < d_U$, test is inconclusive.

③ $H_0: \rho = 0$ Vs. $H_1: \rho \neq 0$

If $d < d_L$ / or $4 - d < d_L$ reject H_0

If $d > d_U$ / or $4 - d > d_U$ accept H_0

Otherwise test is inconclusive.

[For $n = 20$, $d_L = 1.20$, $d_U = 1.41$ ← [Table] $\alpha = 0.05$

$\hat{y}_t = 1608.50 + 20.091\, x_t$

Compute: e_t.

$H_0: \rho = 0$ Vs. $H_1: \rho > 0$

$d = \sum_{i=2}^{20} (e_i - e_{i-1})^2 \Big/ \sum_{i=1}^{20} e_i^2 = 1.08 < d_L = 1.20$,

we reject H_0 and conclude that the errors are positively autocorrelated.]

[For SOFT DRINT CONCENTRATE DATA]

APPLICATION OF LEAST SQUARES REGRESSION TO RELATIONSHIPS CONTAINING AUTO-CORRELATED ERRORS: (JASA '1948)

Cochrane and Orcutt (JASA, 1948) proposed a method to estimate the regression coefficients in the presence of auto-correlated errors.

Consider the simple linear regression model with first order autoregressive errors.

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t \quad \text{where } \epsilon_t = \rho \epsilon_{t-1} + z_t$$
$$z_t \overset{ind.}{\sim} N(0, \sigma_z^2)$$

ρ is an autoregressive parameter.

[NOTE: We can't apply OLS method here since $\epsilon_i \not\overset{ind}{\sim} N(0, \sigma^2)$.]

We transform the response variable $y_t \to y_t' = y_t - \rho y_{t-1}$.

$$y_t' = y_t - \rho y_{t-1} = (\beta_0 + \beta_1 x_t + \epsilon_t) - \rho(\beta_0 + \beta_1 x_{t-1} + \epsilon_{t-1})$$
$$= \beta_0(1-\rho) + \beta_1(x_t - \rho x_{t-1}) + \epsilon_t - \rho \epsilon_{t-1}$$
$$= \beta_0' + \beta_1 x_t' + z_t \quad \text{---(*)}$$

[assuming ϵ_t is first-order autoregressive error]

Now, error z_t are independent and $x_t' = x_t - \rho x_{t-1}$.

Now, we can use OLS on (y_t', x_t'); but (y_t', x_t') time series data cannot be used directly as $y_t' = y_t - \rho y_{t-1}$ and $x_t' = x_t - \rho x_{t-1}$ involve an unknown parameter ρ. ρ is known as autoregressive parameter. How to estimate ρ (Given time series data: (x_t, y_t))?

- We can obtain an estimate of ρ using OLS regression.

$$\epsilon_t = \rho \epsilon_{t-1} + z_t \; ; \; z_t \overset{ind.}{\sim} N(0, \sigma_z^2)$$

Using the given data, we fit SLR model: $y_t = \beta_0 + \beta_1 x_t + \epsilon_t$ using OLS technique and obtain the residuals e_i.

Regress e_i on e_{i-1}; we fit $e_i = \rho e_{i-1} + z_t$

$$S(\rho) = \sum (e_i - \rho e_{i-1})^2$$

$$\frac{\partial S(\rho)}{\partial \rho} = 0 \Rightarrow \sum_{i=1}^{n} (e_i - \rho e_{i-1}) e_{i-1} = 0$$

$$\Rightarrow \rho = \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=2}^{n} e_{i-1}^2}$$

The LSE of ρ is:

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n-1} e_t^2}$$

[Estimate of ρ]

Using this estimate of ρ, we obtain

$$y_t' = y_t - \hat{\rho} y_{t-1} \quad \text{and} \quad x_t' = x_t - \hat{\rho} x_{t-1}$$

and apply OLS to the tranformed data

$$y_t' = \beta_0' + \beta_1 x_t' + z_t \;, \quad z_t \overset{iid}{\sim} N(0, \sigma_z^2)$$

Fitted model: $\hat{y}_t' = \hat{\beta}_0' + \hat{\beta}_1 x_t'$.

- Now, we use Durbin-Watson test to the residual obtained from the reparameterized model. Data: (y_t', x_t').
- If Durbin-Watson test indicates no autocorrelation in the errors, then no additional analysis is needed.
- However, if Durbin-Watson test indicates there is autocorrelation in the errors, then another iteration is required.
- In time series data, observations (errors) are not independent and to test that we can apply residual plots and Durbin-Watson test.
- If we find that autocorrelation exist in the data, then we apply the parameter estimation method by Cochrane and Orcutt (JASA' 1948).

▨ HOMOSCEDASTICITY

e_i's are uncorrelated with each other and have identical mean and variances. If the assumption of homoscedasticity is not met then there can be autocorrelation problem in the data.

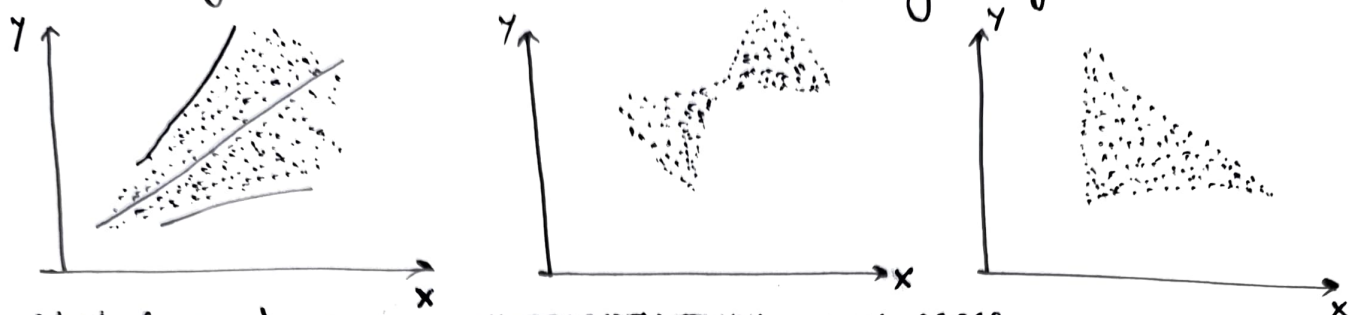
$$Y = X\beta + e$$

$$E(ee') = E((Y - X\beta)(Y - X\beta)') = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \sigma_T^2 \end{pmatrix}$$

If the assumption of homoscedasticity does not hold, then

- OLS estimator, while still unbiased, are no longer providing minimum variance ('BLUE') among the class of linear unbiased estimators.
- Estimated variances of the LS estimators are unbiased, so the usual t and F tests are no longer valid.

Scatters Diagnam of Heteroscedasticity: X: family incomes
Y: family savings



Most frequent case ← HETEROSCEDASTICITY → rare cases

Gleijser Test for Heteroscedasticity:

It regresses the residuals on the explainatory variable ($\hat{Z}$) that is thought to be related to the heteroscedastic variance. Steps are as follows:

Step 1: Fit the OLS regression of Y on X and find the residuals e_i.

Step 2: Regress the absolute value of $|e_i|$ on the explanatory variable that is associated with the heteroscedasticity.

$$|e_i| = \delta_0 + \delta_1 Z_i + u_i$$

$$|e_i| = \delta_0 + \delta_1 \sqrt{Z_i} + u_i$$

$$|e_i| = \delta_0 + \delta_1 \frac{1}{Z_i} + u_i$$

Step 3: Select the equation with the highest R^2 and lowest standard errors to represent heteroscedasticity.

Step 4: $H_0: \delta_1 = 0$ vs. $H_1: \delta_1 \neq 0$ } t-test.

If δ_1 is statistically significant, reject the null hypothesis of homoscedasticity.