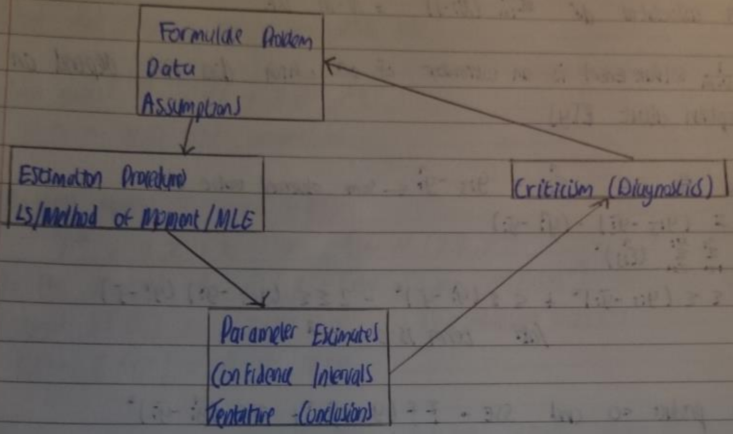
07/12/15 ALSM 1

CHAPTER 3: REGRESSION DIAGNOSTIC



3.1 LACK OF FIT AND PURE ERROR

MSE is an unbiased estimator of σ^2 IF the model is correctly specified, otherwise it is biased

If we repeat observations of y at each value of x (or multiple variables) then we can use these to get an estimate of σ^2 that does not depend on the model

Suppose there are m different values of x: $x_1, \ldots, x_m$ Say at each x_i we measure y N_i times. So there are N_i observations i.e:

$y_{11}\ y_{12} \ldots y_{1N_1} \quad x_1$

$y_{21}, y_{22} \ldots y_{2N_2}$

$\vdots$

$y_{m1}, y_{m2} \ldots y_{mN_m} \quad x_m$

$N = \sum_{i=1}^{m} N_i \qquad$ Let $\bar{y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_j$

Pure error Sum of Squares $SS(\text{Pure Error}) = \sum_{i=1}^{m} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_i)^2$
with associated df $\sum_{i=1}^{m} (N_i - 1) = N - M$ D.F.

So $\frac{1}{N-M} SS(\text{Pure Error})$ is an estimator of σ^2 which does not depend on any assumption about $E[y]$

Consider the SSE $\hat{\varepsilon}_{ij} = y_{ij} - \hat{y}_i$ ← Some observed value at x

$= (y_{ij} - \bar{y}_i) - (\hat{y}_i - \bar{y}_i)$

$= \sum_{i=1}^{m} \sum_{j=1}^{N_i} (\hat{\varepsilon}_{ij})^2$

$= \sum\sum (y_{ij} - \bar{y}_i)^2 + \sum\sum (\hat{y}_i - \bar{y}_i)^2 - 2\sum\sum (y_{ij} - \bar{y}_i)(\hat{y}_i - \bar{y}_i)$

last term is 0

The cross product $= 0$ and $SSE = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum\sum (\hat{y}_i - \bar{y}_i)^2$
$= SS(\text{Pure Error}) + SS(\text{Lack of fit})$

The lack of fit SS has $m-p-1$ d.f. (with p predictors). The test statistic
$F = \dfrac{MS(\text{Lack of Fit})}{MS(\text{Pure Error})}$ follows an $F_{m-p-1,\, n-m}$ d.f. distribution

If error is due to pure error and not just lack of fit.
Suggests that the model is adequate. If we reject, then investigate L.O.F.

3.2 RESIDUAL ANALYSIS

If the model is correct, then error terms have following properties:
zero mean, common variance σ^2, uncorrelated, normal

The residuals $\hat{\varepsilon}_i$ should exhibit similar properties
Hence we often use residual analysis to assess models.
Can be used to measure discrepancy between what we have observed and what we have assumed

14/02/15 ALSM1

3.3 MODEL MISSPECIFICATION

What effect does a misspecified model have?

True model: $y_i = X\beta + Z\gamma + \varepsilon_i$ — additional design matrix

$\varepsilon_i \sim N(0, \sigma^2)$, uncorrelated

Analyst's model: $y_i = X\beta + \varepsilon_i^*$

$\varepsilon_i^* = Z\gamma + \varepsilon_i \Rightarrow \varepsilon_i^* \sim N(Z\gamma, \sigma^2)$

In this situation, $\hat{\beta} \sim N\left(\beta + (X^TX)^{-1}X^TZ\gamma,\ \sigma^2(X^TX)^{-1}\right)$

and $\mathbb{E}[\hat{\varepsilon}^*] = (I-H)Z\gamma$ SEE PROBLEM SHEET 3

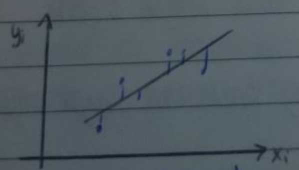
with $\mathbb{E}[\hat{\varepsilon}_i^*] = 0$ if $\gamma = 0$

As an example: $X\beta = \beta_0 + \beta_1 x_i$ (SLR) but we actually should fit a quadratic model

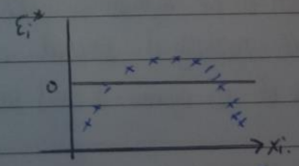
i.e: $Z\gamma = \beta_2 x_i^2$

Then, it can be shown that $\mathbb{E}[\hat{\varepsilon}_i^*] = \beta_2\left[x_i^2 - \frac{\sum x_i^2}{n} - (x_i - \bar{x})\frac{\sum x_i^2(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\right]$

Unless $\beta_2 = 0$, the i^{th} residual is a quadratic function of x_i, hence a plot of the residuals against x_i



Interested in deviations (residuals)



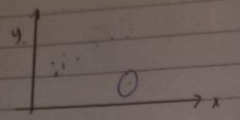
Residuals: Quadratic Shape

Should look curved in appearance

It is often useful to plot $\hat{\varepsilon}^*$ against $\hat{y}_i$

3.4 DETECTION OF OUTLIERS

Outliers are individual observation points which don't follow the same model assumed for the rest of the data.



- The size of the standardised residuals $\frac{\hat{\varepsilon}_i}{\sigma\sqrt{(I-H_{ii})}}$ is an important diagnostic
- If σ is replaced by $\sqrt{MSE}$ we have the studentised residual
- However, if the i^{th} point is an outlier, MSE will be biased upwards (Bigger than it should be)
- Instead we use the MSE from a fit with the i^{th} observation deleted: MSE_{-i}
- "Internal Studentised Residual" $r_i = \hat{\varepsilon}_i / \sqrt{MSE\,(I-H_{ii})}$ (i^{th} point included)
- "External Studentised Residual" $t_i = \hat{\varepsilon}_i / \sqrt{MSE_{-i}\,(I-H_{ii})}$ (i^{th} excluded)

  if point outside ± 2 then it is most likely an outlier

3.5 CHECKING FOR NORMALITY

- Inference (CIs, hypothesis test) require a normal assumption
- Minor departures from normality are insignificant
- Model misspecification could lead to false detection of departure from normality

Normal Probability Plot

- Order the studentised residuals $r_1 \leq r_2 \leq \ \leq r_n$
- These are the sample order statistics. If these are $N(0,1)$ then $\mathbb{E}[r_{(i)}] = \mathbb{E}[z_{(i)}]$

  which is the expected value for the i^{th} order statistic from a $N(0,1)$
- So, if normality is satisfied, $r_{(i)} = \mathbb{E}[z_{(i)}] + \text{error}$.
- A plot of $r_{(i)}$ against $\mathbb{E}[z_{(i)}]$ should be approximately linear ($45°$ through origin).

3.6 INHOMOGENEOUS VARIANCE

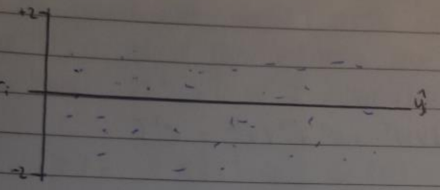
- There is a problem in using the ordinary residuals $\hat{\varepsilon}_i$ as a diagnostic
- Consider the vector $\hat{\underline{\varepsilon}}$ where $\hat{\underline{\varepsilon}} = \underline{y} - \hat{\underline{y}} = \underline{y} - H\underline{y} = (I-H)\underline{y}$ $\quad H = (X^TX)^{-1}X^T$
- $Var[\hat{\underline{\varepsilon}}] = (I-H)\,Var[\underline{y}]\,(I-H)^T$
- $Var[\underline{y}] = \sigma^2 I$ (Assumption in model)
- $Var[\hat{\underline{\varepsilon}}] = \sigma^2 (I-H)(I-H)^T = \sigma^2 (I-H)$ While $Var[\underline{\varepsilon}] = \sigma^2 I$ ← no hat
- So $Var[\hat{\varepsilon}_i] = \sigma^2 (I - H_{ii})$

14/12/15 ALSM 1

- H_{ii} is the i^{th} diagonal element of H. If there is large variation in the diagonal element of H, there will be large differences in the variances of the residuals.
- As a result, we work with standardised residuals: $\frac{\hat{\varepsilon}_i}{\sigma\sqrt{(I-H_{ii})}}$

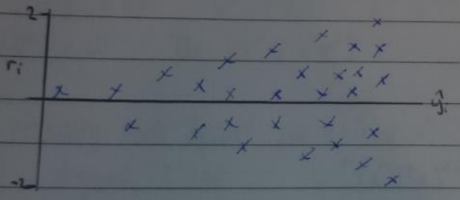
Studentised Residual $= \hat{\varepsilon}_i / (\sqrt{MSE}\sqrt{(I-H_{ii})})$

Homogeneous Variance:



- no obvious pattern
- scattered between ±2
- $r_i \Rightarrow$ studentised residuals

Inhomogeneous Variance (depends on predictor):



- Fans outwards or inwards

3.7 NON-STANDARD CONDITIONS - TRANSFORMATIONS OF VARIABLES

- transformation to stabalise variance
- transformation to achieve normality

Variance Stabalising Transformations

- Consider $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$ with $E[\underline{\varepsilon}] = \underline{0}$ $Var[\underline{\varepsilon}] = \sigma^2 V$ ← diagonal matrix where $V \neq I$
- Weighted least squares assumes errors are uncorrelated but heteroskedastic errors.

$$V = \begin{bmatrix} a_1^2 & & 0 \\ & a_2^2 & \\ 0 & & \ddots & \\ & & & a_n^2 \end{bmatrix}$$ ← a's not σ's

- So that $Var[\varepsilon_i] = a_i^2\sigma^2$ $Cov[\varepsilon_i, \varepsilon_j] = 0$ $i \neq j$
- Consider the transformation $z_i = y_i / a_i$

- $Var[z_i] = \frac{1}{a_i^2} Var[y_i] = \frac{a_i^2 \sigma^2}{a_i^2} = \sigma^2$
- $\underline{z}$ is homoscedastic
- Let $W = [V^{1/2}]^{-1}$ inverse of root

$$W = \begin{bmatrix} 1/a_1 & & 0 \\ & 1/a_2 & \\ 0 & \ddots & 1/a_n \end{bmatrix}$$

$z = W\underline{y} \qquad \underline{y} = X\beta + \underline{\varepsilon}$

$z = WX\beta + W\underline{\varepsilon}$

$\underline{z}^* = X^*\beta + \underline{\varepsilon}^*$

- Where $X^* = WX \qquad \underline{\varepsilon}^* = W\underline{\varepsilon}$
- Then $\mathbb{E}[\underline{\varepsilon}^*] = 0 \quad Var[\underline{\varepsilon}^*] = \sigma^2 WVW^T = \sigma^2 V^{-1/2} V V^{-1/2} = \sigma^2 I$

- Ordinary least squares can be used on $\underline{z}$ and X^*
- The weighted least squares estimator is:

$$\hat{\beta} = (X^{*T}X^*)X^{*T}\underline{z}$$
$$= (X^T W^T W X)^{-1} X^T W^T W \underline{y}$$
$$= (X^T V^{-1} X)^{-1} X^T V^{-1} \underline{y}$$

$\mathbb{E}[\hat{\beta}] = \beta$

$Var[\hat{\beta}] = \sigma^2 (X^{*T}X^*)^{-1} = \sigma^2 (X^T V^{-1} X)^{-1}$

Example:

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad Var[\varepsilon_i] = a_i^2 \sigma^2$

$z_i = y_i/a_i = \beta_0/a_i + \beta_1 x_i/a_i + \varepsilon_i/a_i$

$\mathbb{E}[z_i] = \beta_0/a_i + \beta_1 x_i/a_i$

$SSE = \sum_{i=1}^{N} (z_i - \hat{z}_i)^2 = \sum_{i=1}^{n} (y_i/a_i - \hat{\beta}_0/a_i - \hat{\beta}_1 x_i/a_i)^2$

$= \sum_{i=1}^{N} 1/a_i^2 \, (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

$= \sum \frac{1}{a_i^2} (y_i - \hat{y}_i)^2$

weights $\Rightarrow$ contribution for each i

Consider when a_i is small/large and how reliable collected data are in these cases

14/12/15 ALSM I 5

y_i is large variance → unreliable → small weight
y_i is small variance → reliable → large weight

Let $w_i = 1/a_i^2$ be the weights
$SSE = \sum w_i (y_i - \hat{y}_i)^2$ Weighted least squares WLS

3.8 GENERALISED LEAST SQUARES

- In the general case $Var[\varepsilon] = \sigma^2 V$ where V is not necessarily diagonal
- Since V is positive definite matrix, there exist a $n \times n$ non singular matrix T, such that $TT^T = V$ T - called cholesky triangle

$\underline{y} = X\beta + \varepsilon$ $\quad E[\varepsilon] = 0 \quad Var[\varepsilon] = \sigma^2 V$

$\underline{z} = T^{-1}\underline{y} = T^{-1}(X\beta + \varepsilon)$
$= T^{-1}X\beta + T^{-1}\varepsilon$

$\underline{z} = X^*\beta + \varepsilon^* \quad X^* = T^{-1}X \quad \varepsilon^* = T^{-1}\varepsilon$

$E[\varepsilon^*] = 0 \quad Var[\varepsilon^*] = \sigma^2 I$

- Use ordinary least squares on $\underline{z}$ and X^*
- If V is diagonal, $T^{-1} = W$
- WLS is a special case of generalised least squares.

3.9 VARIABLE SELECTION

- Have candidate regressors - want subset to use in model
- If we use all the regressors:
  - Unbiased estimates
  - Large variance of parameter estimates and predicted value
  - More costly (more calculations)
  - Cost of inverting $n \times n$ symmetric matrix scales cubically with n (cholesky decomposition)

- If we use a subset:
  - Reduced cost
  - Possibly biased parameter estimates
  - Reduced variance of the parameter estimates and predicted values

How difficult to chose a subset?

- If p possible predictors/regressors
  $\binom{p}{0} + \binom{p}{1} + \dots + \binom{p}{p} = 2^p$ possible models
- If p is large (wide data), this is alot.

Forward Selection

- Begin with $y_i = \beta_0 + \beta_K x_{Ki} + \varepsilon_i$ where x_K is the X which gives the largest R^2 on its own
- Then add x_J to the model s.t. $y_i = \beta_0 + \beta_K x_{Ki} + \beta_J x_{Ji} + \varepsilon_i$ x_J such that the largest increase in R^2 is achieved
- Repeat until some stopping criterion is satisfied e.g. the F-test for each of the variables that not yet entered is less than some pre-determined value

Stepwise Regression

- Begin as for forward selection, then at each step remove one of the variables in the current model if has $F <$ predetermined value (partial F-test)
- Similarly add a variable not included in the model if $F >$ pre-determined value
- Iterate under no further additions or removals

All Possible Regressions

- Fit all possible models.
- Only an option if you have a small number of variables (2^p models)
- We can consider some statistic for each e.g. (likelihood criterion, AIC, BIC)
- Max likelihood - penalty function - depends on # variables in model - penalises complex models.
- Lots of possible statistics e.g. Mallows suggested statistic for model with p predictors:
  $$C_p = \frac{SSE(p)}{\hat{\sigma}^2} - n + 2p$$
- One can draw parallels with AIC & BIC. Can show $E[C_p] = p$