

**Criteria for the interpretation of selected statistics from the SAS output**

**A) General regression diagnostics**

1) **Adjusted  $R^2$**  :  $R_{adj}^2 = 1 - \frac{(n-1)}{(n-p)} \left( \frac{SSE_{Error}}{SST_{Total}} \right) = \left( \frac{n-1}{n-p} \right) (1 - R^2)$

- a) This is intended to be an adjustment to  $R^2$  for additional variables in the model
- b) Unlike the usual  $R^2$ , this value can decrease as more variables are entered in the model if the variables do not account for sufficient additional variation (equal to the MSE).

2) **Standardized regression coefficient  $b_j'$**  .  $b_j' = (S_{x_j} / S_y)$

- a) Unlike the usual regression coefficient, the magnitude of the standardized coefficient provides a meaningful comparison among the regression coefficients.
- b) Larger standardized regression coefficients have more impact on the calculation of the predicted value and are more "important".

3) **Partial correlations**

- a) Squared semi-partial correlation TYPE I = SCORR1 =  $SeqSSX_j / SST_{Total}$
- b) Squared partial correlation TYPE I = PCORR1 =  $SeqSSX_j / (SeqSSX_j + SSE_{Error})$
- c) Squared semi-partial correlation TYPE II = SCORR2 =  $PartialSSX_j / SST_{Total}$
- d) Squared partial correlation TYPE II = PCORR2 =  $PartialSSX_j / (PartialSSX_j + SSE_{Error})$
- e) Note that for regression, TYPE II SS and TYPE III SS are the same.

**B) Residual Diagnostics**

- 1) The hat matrix main diagonal elements,  $h_{ii}$  (Hat Diag , H values in SAS) , called "leverage values", they are used to detect outliers in X space. . This can also identify substantial extrapolation of new values. As a general rule,  $h_{ii}$  values greater than 0.5 are "large" while those between 0.2 and 0.5 are moderately large, also look for a leverage value which is noticeably larger than the next largest

- a) The  $h_{ii}$  values sum to p mean,  $\bar{h}_{ii} = p/n$  (note that this is  $< 1$ )
- b) A value may be an outlier if it is more than twice the value  $\bar{h}_{ii}$  (i.e.  $\bar{h}_{ii} > 2p/n$ ).

- 2) Studentized residuals ("Student Residual" in SAS). Also called Internally Studentized Residual.

- a) There are two versions:

Simpler calculation =  $e_i / \text{root}(MSE)$

More common application =  $e_i / \text{root}(MSE * (1-h_{ii}))$  [SAS produces these

- b) We already assume these are normally distributed, so these values would approximately follow a t distribution, where for large samples

about 65% are between -1 and +1

about 95% are between -2 and +2

about 99% are between -2.6 and +2.6

- 3) Deleted Studentized residuals ("RStudent" in SAS). Also called externally studentized residual.

- a) There are also two versions as with the studentized residuals above

Deleted Studentized =  $e_{i(i)} / \text{root}(MSE_{(i)})$

Deleted Internally Studentized =  $e_{i(i)} / \text{root}(MSE_{(i)} * 1-h_{ii})$  [values produced by SAS]

- b) As with the studentized residuals above these values would approximately follow a t distribution

**C) Influence Diagnostics**

- 1) DFFITS; an influence statistic, it measures the difference in fits as judged by the change in predicted value when the point is omitted

- a) This is a standardized value and can be interpreted as the number of standard deviation units

- b) for small to medium size databases, DFFITS should not exceed 1, while for large databases it should not exceed  $2\sqrt{p/n}$
  - 2) DFBETAS; an influence statistic, it measures the difference in fits as judged by the change in the values of the regression coefficients
    - a) note that this is also a standardized value
    - b) for small to medium size databases, DFBETAS should not exceed 1, while for large databases it should not exceed  $2/\sqrt{n}$
  - 3) Cook's D : influence statistic (D is for distance)
    - a) The boundary of a simultaneous regional confidence region for all regression coefficients
    - b) this does not follow an F distribution, but it is useful to compare it to the percentiles of the F distribution  $[F_{1-\alpha; p, n-p}]$  where a change of  $< 10^{\text{th}}$  or  $20^{\text{th}}$  percentile shows little effect, while the  $50^{\text{th}}$  percentile is considered large
- D) multicollinearity Diagnostics
- 1) VIF is related to the severity of multicollinearity
    - a) a standardized estimate of regression coefficients would be expected to have a value of 1 if the regressors are uncorrelated
    - b) If the mean of this value is much greater than 1, serious problems are indicated.
    - c) No single VIF should exceed 10
  - 2) Tolerance is the inverse of VIF, where  $\text{Tolerance}_k = 1 - R_k^2$
  - 3) The Condition number (a multivariate evaluation)
    - a) Eigen values are extracted from the regressors, These are variances of linear combinations of the regressors, and go from larger to smaller.
    - b) If one or more are zero (at the end) then the matrix is not full rank.
    - c) These sum to p, and if the  $X_k$  are independent, each would equal 1
    - d) The condition number is the square root of the ratio of the largest (always the first) to each of the others.
    - e) If this value exceeds 30 then multicollinearity may be a problem.
- E) Model Evaluation and Validation
- 1)  $R_p^2$ ,  $\text{Adj}R_p^2$  and  $\text{MSE}_p$  can be used to graphically compare and evaluate models. The subscript p refers to the number of parameters in the model
  - 2) Mallows's  $C_p$  criterion
    - a) Use of this statistic presumes no bias in the full model MSE, so the full model should be carefully chosen to have little or no multicollinearity
    - b)  $C_p$  criterion =  $(\text{SSE}_p / \text{TrueMSE}) - (n - 2p)$
    - c) The  $C_p$  statistics will be approximately equal to p if there is no bias in the regression model
  - 3)  $\text{PRESS}_p$  criterion ( $\text{PRESS} = \text{Prediction SS}$ )
    - a) This criterion is based on deleted residuals.
    - b) There are n deleted residuals in each regression, and  $\text{PRESS}_p$  is the SS of deleted residuals
    - c) This value should approximately equal the MSE if predictions are good, it will get larger as predictions are poorer
    - d) They may be plotted, and the smaller PRESS statistic models represent better predictive models.
    - e) This statistics can also be used for model validation

Reference was primarily Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W., Applied Linear Statistical Models, 4th Edition, Richard D. Irwin, Inc., Burr Ridge, Illinois, 1996.