Building the Regression model: Diagnostic

1. Model adequacy for a predictor variable

* limitation of residuals obtained from regression

$$Y = Xb + \varepsilon$$

is that they likely do not properly show the nature of the marginal effect of a predictor variable, given the other predictors in the model.

To identify the nature of the marginal relationship for a predictor variable X_k we can use the so called Partial Regression plots.

PRP should be use with care, as they may provide erroneous information (or inference based upon) if the relationships of some predictors kept in the model are misspecified.

For example if X_2 & X_3 are nonlinear related but in the model they are linearly combined then PRP of X_2 & X_3 could be misleading.

The way how PRP are obtained is somehow indirect because we want to plot $e(Y|X_k)$ against $e(X_k|X_1, X_2, \ldots, X_{k-1}, X_{k+1}, \ldots X_p)$.

Therefore, we have to compute the two error terms which are obtained from two regressions:

$$Y = X_k b + e(Y|X_k)$$

$$X_k = X' b + e(X_k|X')$$

where $X' = (X_1, X_2, \ldots X_{k-1}, X_{k+1}, X_p)$

2. Identifying Outlying Observation in PREDICTED variable

Frequently dataset contains measurements that are very different from the rest, which are called outliers.

Outliers have dramatic effects particularly for OLS. Therefore we have to identify them and assess them formaly.

An easy way to see if there are outliers in the data is thru Box-Plots.

However outliers identified with Box-plots are only for visual assement.

For formal assessment we will be using the following statistics :

1) variance of residual of observation i.
2) studentized deleted residuals.

To estimate these two statistic lets remember that

a) $e_i = Y_i - \hat{Y}_i$ and the semistudentized residuals $e_i^* = \frac{e_i}{\sqrt{MSE}}$

b) Hat matrix $H = X(X^tX)^{-1}X^t$.

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

$$\sigma^2(e_i) = \sigma^2(1 - h_{ii})$$

$$s^2(e_i) = MSE(1 - h_{ii})$$

Now that we remember the notation we can compute further statistics:

a) studentized residuals

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

b) Deleted residual

Remember PRESS statistic: $y_i - \hat{y}_{i(i)}$

Let define $d_i = y_i - \hat{y}_{i(i)}$ being the residual of the measured y_i and the predicted $\hat{y}_i$ but the estimates computed without the i^{th} observation

d_i can be estimated without refitting the regressions.

$$d_i = \frac{e_i}{1-h_{ii}}$$

$$s^2(d_i) = MSE_{(i)}\left(1 + x_i^t (X_{(i)}^t X_{(i)})^{-1} x_i\right)$$

An algebraically equivalent of $s^2(d_i)$ is.

$$s^2(d_i) = \frac{MSE_{(i)}}{1-h_{ii}}$$

$$\frac{d_i}{s(d_i)} \sim t(n-p-1).$$

c). Studentized Deleted Residuals.

$$t_i = \frac{d_i}{s(d_i)}$$

An algebraically equivalent expression is.

$$t_i = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

However, the studentized deleted residuals can be computed without refitting a new regression each time an observation is omitted.

$$MSE_{(i)} = \frac{n-p}{n-p-1} MSE - \frac{e_i^2}{(1-h_{ii})(n-p-1)}$$

Therefore, $t_i = e_i \sqrt{\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}}$

$t_i \sim t(n-p-1)$

Note Observation

The assessment is executed using the multiple comparisons test whose critical values are $t(1-\frac{\alpha}{2n}; n-p-1)$ which are $>> t(1-\frac{\alpha}{2}, n-p-1)$

This means that p-values very small can be, indicate outliers.

3. Identifying Outlying observations in PREDICTOR variables

Hat matrix plays an important role not only on the predictor but also on the predicter outliers.

H matrix has some useful properties.

1) $h_{ii} \geq 0$
2) $h_{ii} \leq 1$
3) $\Sigma h_{ii} = p$. (trace of H is p)
4) h_{ii} is a measure of the distance between the X values for the i^{th} observation and the mean of X. Therefore, a large h_{ii} indicate that the i^{th} observation is far from the center of all X observations

In this context the diagonal element h_{ii} is called Leverage.

Notes. $H = X^t (X^t X)^{-1} X$ which means h_{ii} depends only on X.

A leverage value h_{ii} is considered large if it is twice (2) as large than mean leverage.

$$\bar{h} = \frac{\Sigma h_{ii}}{n} = \frac{p}{n}.$$

Therefore $h_{ii} > 2\bar{h}$ are outlying $\Rightarrow$ $\boxed{h_{ii} \leq \frac{2p}{n}}$

4. Identifying Influential Observations.

Not all outliers are influential.

An influential observation is an observation whose exclusion causes major changes in fitted regression

a) Influence on single fitted value
b) Influence on all fitted values
c) Influence on regression coefficients

a) Influence on single fitted value

The statistics: $(DFFITS)_i = \dfrac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} \cdot h_{ii}}}$

DFFITS = difference (DF) between fitted (FIT) value $\hat{Y}_i$ when all observations (S) are used and the fitted value of $\hat{Y}_i$ when the i^{th} case is omitted.

$$(DFFITS)_i = e_i \cdot \sqrt{\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}\left(\frac{h_{ii}}{1-h_{ii}}\right)} = t_i \sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

where t_i is the studentized deleted residual.

RULE OF THUMB: an observation is influential if

$|DFITS| > 1$ if n small

$|DFITS| > 2 \cdot \sqrt{\frac{p}{n}}$ if n large

b) Influence on all fitted values. = Cook's distance.

Cook's distance is the opposite of DFITS, in the sense that it considers the influence of the i^{th} observation on the other observations.

The Cook's distance is: $D_i = \frac{\sum (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \cdot MSE}$

Cook's distance can be calculated from original regression.

$$D_i = \frac{e_i}{p MSE} \left[\frac{h_{ii}}{(1-h_{ii})^2} \right]$$

D_i is related to F distribution percentiles.

Guideline: if $P(F_{p,n-p} \leq D_i) \leq 0.2$ then non-influential

if $P(F_{p,n-p} \leq D_i) > 0.5$ the influential

c) Influence on the regression coefficients:

A measure of influence of the i^{th} observation on each regression coefficient b_k is the difference between the estimated b_k with all observations and the same $b_{k(i)}$ with the i^{th} observation omitted, $b_{k(i)}$.

$$DFBETAS = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}}$$

where c_{kk} = the k^{th} element of the $(X^t X)^{-1}$ matrix.

$$\sigma^2(b) = \sigma^2 (X^t X)^{-1}$$

$$\sigma^2(b_k) = \sigma^2 \cdot c_{kk}$$

Guideline large absolute value of $DFBETAS_{k(i)}$ indicate large impact of i^{th} obs. on k^{th} coefficient

if $|DFBETAS| > 1$ influential for n small

$|DFBETAS| > \frac{2}{\sqrt{n}}$ influential for n-large.

Comments

1) Analysis of outlying and influential observations is MANDATORY in Regression analysis.

2). Detection of outliers and influential observations can be ineffective.

An example is 2 outliers close by where only one is initialy identified as outliers, the second one being "masked" by the first one.

Multicolinearity

Remember that the inverse of a matrix exists only if the determinant is non-zero, meaning the matrix is non-singular.

A singular matrix is a matrix with determinant 0 or that have linear dependencies among rows or columns.

When predictor variables are correlated we say that they are MULTICOLINEAR, particularly if the correlation is high.

let take an example.

X_1	X_2	Y
2	6	23
8	9	83
6	8	63
10	10	103

$$\begin{cases} \hat{y} = -87 + X_1 + 18X_2 \\ \hat{y} = -7 + 9X_1 + 2X_2 \end{cases}$$

Different functions for the same data, is something that we do not want

This happens because there is a perfect relationship between X_1 & X_2:

$$X_2 = 5 + \frac{1}{2} X_1$$

Observations

1) Multicolinearity does not preclude a good model
2) we cannot decide which variable is more important based on model as X_1 has $b_1 > b_2$ in first model and X_2 has $b_2 > b_1$ in second model.

Effects of multicolinearity.

1) on Regression coefficients
2) on Sum of Squares.
3) on $S[b_k]$
4) on fitted values
5) effects on Simultaneous Test of β_k.

Proof with example.

Body fat examples

1) Regression coefficients:

Model	b_1	b_2	b_3
Triceps	0.85 (0.13)		
Thigh	0.85 (0.11)	0.85 (0.11)	
Midarm			0.199 (0.32)
Triceps + Thigh	0.22 (0.30)	0.66 (0.29)	
Triceps + Midarm	1.0 (0.13)		-0.43 (0.17)
Thigh + Midarm		0.85 (0.11)	0.09 (0.16)
Triceps + Thigh + Midarm	4.3 (3.0)	-2.8 (2.5)	-2.1 (1.6)

3). effects on $s(b_k)$
see table before

Multicolinearity Diagnostic

A. Informal diagnostics

a). large change in regression coefficients when variables are added or deleted

b). nonsignificant results on regression coefficients for important predictor variables

c). coefficients have different signs from expected experience or intuition

d) large correlation coefficients between pairs of predictor variables

e) wide confidence intervals for regression coefficients

B. Formal diagnostic = Variance Inflation Factor

VIF measure how much variance of the estimated coefficients are inflated compared to when the predictors are not linearly related.

we know that

$$\sigma^2[b] = \sigma^2(X^tX)^{-1}$$

To measure impact on colinearity alone it is important to eliminate other aspects of data, such as magnitude.

Therefore, VIF are computed on standardized regression.

$$y'_i = \frac{1}{\sqrt{n-1}} \left(\frac{y_i - \bar{y}}{s_y} \right)$$

$$x'_{ik} = \frac{1}{\sqrt{n-1}} \left(\frac{x_{ik} - \bar{x}_k}{s_k} \right)$$

The standardized regression model does not contains an intercept term.

Therefore $X' = \begin{pmatrix} x'_{11} & \cdots & x'_{1,p-1} \\ \vdots & & \\ x'_{n1} & \cdots & x'_{n,p-1} \end{pmatrix}$

Let define $r_{xx} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1,p-1} \\ r_{21} & & & \\ \vdots & & & \\ r_{p-1,1} & r_{p-1,2} & \cdots & r_{p-1,p-1} \end{pmatrix}$

The standardized regression is:

$$y' = X' b' + \varepsilon$$

and the parameters of the original regression are

$$b_k = \left(\frac{s_y}{s_k} \right) b'_k$$

It is a simple exercise to show that

$$\sigma^2(b) = (\sigma')^2 r_{xx}^{-1}$$

$(\sigma')^2$ is the variance of of the standardized model.

$$\sigma^2(b'_k) = (\sigma')^2 \cdot (r_{xx}^{-1})_k \Rightarrow$$

where $(r_{xx}^{-1})_k$ is the k^{th} diagonal element of r_{xx}^{-1}

If we define $VIF_k = (r_{xx})^{-1}_{kk}$ then

$$VIF_k = (1 - R_k^2)^{-1}$$

where R_k^2 is the coef. of determination when X_k is regressed against the $p-2$ other X

$$X_k = b_0 + b_1 X_1 + \dots + b_{k-1} X_{k-1} + b_{k+1} X_{k+1} + \dots + b_p X_p$$

Therefore $\sigma^2(b'_k) = \dfrac{(\sigma')^2}{1 - R_k^2}$

Observation: R_k^2 increases $\Rightarrow$ VIF_k increases.

Guideline: If $VIF > 10$ then multicolinearity is present.

Refresh assumptions.

1) linearity
2) No multicolinearity
3) $\varepsilon \sim N(0, \sigma^2)$
4) residuals are non-correlated.
5) homoskedasticity

Diagnostic = identify problems:

→ computational: outliers
: singularities.

→ model fit : linearity.

→ prediction : → normality
→ homoskedasticity
→ non-correlation

Up to now we focused on computational and model fit issues:

Next we will spend time on identifying issues in prediction:

a) Normality:

Shapiro-Wilks test ⇒ based on regression

Anderson-Darling test ⇒ based on EDF

(in the same family with KS/Kuiper/Cramer-von Mises).

b) Homoskedasticity:

1) Breush-Pagan test.

Assume that error terms are independent and normal. Also assume that variance is related to level of X.

$\sigma_i^2 = e^{\gamma_0 + \gamma_1 X_i} \Rightarrow \log \sigma_i^2 = \gamma_0 + \gamma_1 X_i$

$H_0: \gamma_1 = 0 \Rightarrow \chi^2_{BP} = \frac{SSR^*}{2} / \left(\frac{SSE}{n}\right)^2$

SSR^* - SSR when regressing e^2 against X.

2) modified Levene test - does not depend on normality.

c) Non-correlation

Durbin-Watson test.

DW is a first-order autoregressive error model.

$e_i = Y_i - \hat{Y}_i$

Define the statistics: $D = \frac{\sum (e_i - e_{i-1})^2}{\sum e_i^2}$

Durbin-Watson bounds for D: d_u & d_L

If $D > d_u$ ⇒ no auto correlation

$D < d_L$ ⇒ Auto correlation

$d_L < D < d_u$ ⇒ inconclusive