MA 542 SPRING 2018

Applied Regression Analysis

Chapter 10

Building the Regression Model II: Diagnostics

Diagnostics

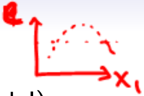
In this chapter we discuss number of refined diagnostics for checking the adequacy of a regression model. These include methods for detecting

$SSR(X_1|X_2)$ $R^2_{Y1|2}$

- Improper functional form for a predictor variable,
- Outliers,
- Influential observations,
- Multicollinearity.

.

Limitations of Residual Plots



- Residual plots vs. the predictor variables (in the model) can be used to check whether curvature effect for that variable is required in the model.

- Residual plots vs. the predictor variables (not yet in the model) can be used to determine adding one or more of these variables to the model.

- **Limitation:** These plots do not show the nature of the marginal effect of a predictor variable, given the other predictor variables in the model.

Added Variable Plots

Added Variable Plots (partial regression plots or adjusted variable plots) provide graphic information about the marginal importance of a predictor variable X_k , given the other predictor variables already in the model.

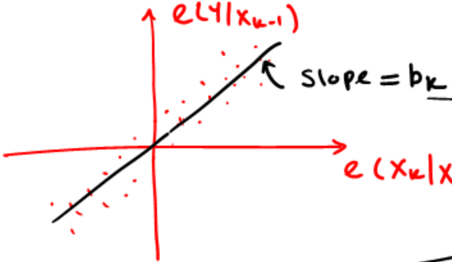
Here is how we draw the plot: X_k

- Regress Y vs. all predictors except X_k, calculate the residuals $e(Y|X_{-k})$. $(Y \sim X_1, X_2, \ldots X_{k-1}, X_{k+1}, \cdots, X_{p-1})$
- Regress X_k vs. all predictors except X_k, calculate the residuals $e(X_k|X_{-k})$. $(X_k \sim X_1, X_2, \ldots X_{k-1}, X_{k+1}, \ldots X_{p-1})$
- Then the *added variable plot* is the plot of $e(Y|X_{-k})$ vs $e(X_k|X_{-k})$.

Added Variable Plots

The least squares estimate b_k obtained from fitting a line (through the origin) to the added variable plot is same as one would get from fitting the full model

$$Y = \beta_0 + \beta_1 X_{i1} + \cdots \beta_k X_{ik} + \cdots + \beta_{p-1} X_{ip-1} + \epsilon \text{ } (\textit{Christensen}, \text{ 1996}).$$

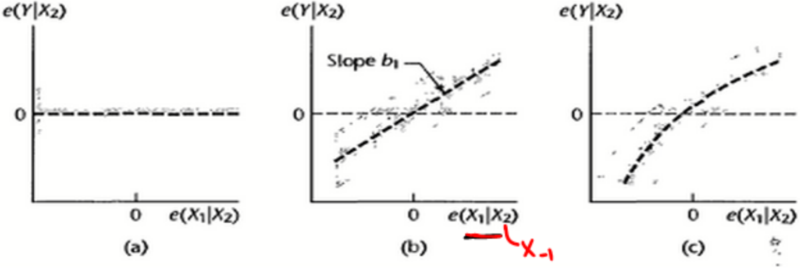


Fitted model:

$$Y = b_0 + b_1 X_{i1} + \cdots b_k X_{ik} + \cdots + b_{p-1} X_{ip-1}.$$

$\textcircled{X_1}$

Prototype Added Variable Plots

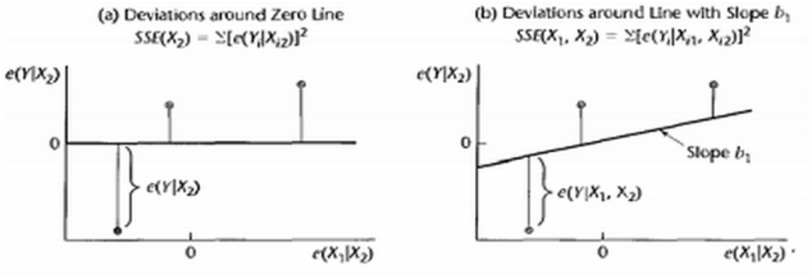


(a) A horizontal band: X_1 contains no additional information; adding X_1 is not suggested.

(b) A linear band with a nonzero slope; X1 may be a helpful addition to the regression model already containing X2.

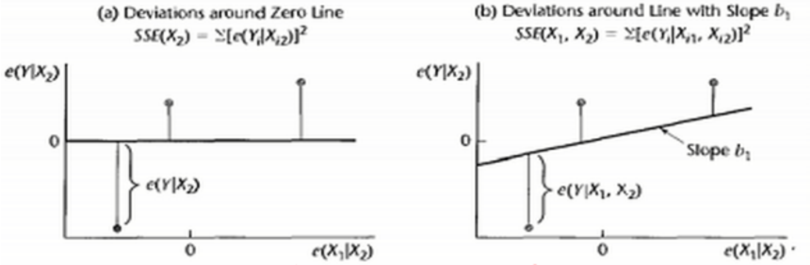
(c) A curvilinear band; X1 may be helpful and suggesting the possible nature of the curvature effect by pattern shown.

Strength of the Linear Relation



- $SSE(X_2)$: Sum of the square deviations from the zero line.
- $SSE(X_1, X_2)$: Sum of the square deviations from the line with slope b_1.
- $SSR(X_1|X_2) = SSE(X_2) - SSE(X_1, X_2)$: Strength of the linear relation of X1 to the response variable, given that X2 is in the model.

Strength of the Linear Relation



$SSR(X_1|X_2) = SSE(X_2) - SSE(X_1, X_2)$

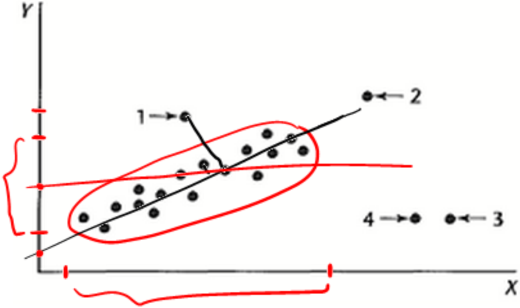
- If all the points are close to the line with slope b_1, then $SSE(X_1, X_2) <<< SSE(X_2)$. So $SSR(X_1|X_2)$ is large and X_1 should be included in the model.

- If all the points are close to the horizontal line, then $SSE(X_1, X_2) \approxeq SSE(X_2)$. So $SSR(X_1|X_2)$ is small and X_1 is not that important in the model.

Outlying or Extreme Observations

- The observations (cases) that are well separated from the remainder of the data are called outliers or extreme cases.
- Outlying cases may involve large residuals.
- A case may be outlying with respect to its Y value, its X value, or both.
- Not all outlying cases have a strong influence on the fitted regression function.
- Key step: Determining whether the regression model under consideration is heavily influenced by one or a few cases in the data set.

Outlying or Extreme Observations (E.g.)



- Case 1 :
  Outlying with respect to Y : may not be too influential.

- Case 2 :
  Outlying with respect to both : may not be too influential.

- Cases 3 & 4 :
  Outlying with respect to X : likely to be very influential.

Identifying Outliers with respect to Y

$\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$

1. Residuals, Semistudentized Residuals:

$$e_i = Y_i - \hat{Y}_i, \qquad e_i^* = \frac{e_i}{\sqrt{MSE}}$$

2. Hat matrix: $H = X(X'X)^{-1}X' \Rightarrow \hat{Y} = HY$.

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} \cdots X_{1\,p-1} \\ 1 & X_{21} & X_{22} \cdots X_{2\,p-1} \\ \vdots & \vdots & \\ \cdot & \cdot & \cdots \end{bmatrix}$$

Residual vector: $e = (I - H)Y$.

$\sigma^2\{Y\} = \sigma^2$

$\mathrm{Var}(AX) = A\underbrace{\mathrm{Var}(X)}_{\sigma^2}A^T = \sigma^2 \underbrace{AA^T}_{A^2}$

$\sigma^2\{e\} = \sigma^2(I - H)$

$\Rightarrow \sigma^2\{e_i\} = \sigma^2(1 - h_{ii})$, and $\sigma\{e_i, e_j\} = -h_{ij}\sigma^2, \quad i \neq j,$

where h_{ii} : the i^{th} element of the diagonal of H and,

h_{ij} : the ij^{th} element of the matrix H.

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & & & \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix}$$

Then the estimated variance,

$$\Rightarrow s^2\{e_i\} = MSE(1 - h_{ii}), \text{and } s\{e_i, e_j\} = -h_{ij}MSE, \quad i \neq j,$$

Note: $h_{ii} = X_i'(X'X)^{-1}X_i, \qquad X_i = [1, X_{i,1}, X_{i,2}, \cdots, X_{i,p-1}]'$.

Deleted Residuals

- The difference between Y_i and $\hat{Y}_{i(i)}$ (fitted value of the model with all but i^{th} case) is called the i^{th} deleted residual.

$$d_i = Y_i - \hat{Y}_{i(i)}$$

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

- It can be shown that $d_i = Y_i - \hat{Y}_{i(i)} = \boxed{\frac{e_i}{1-h_{ii}}}$ – using the full data set
- The estimated variance of d_i :

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

- It also can be shown that

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

.

Studentized Deleted Residuals

- $t_i = \frac{d_i}{s\{d_i\}}$ $\sim t(n-1-p)$ is called the i^{th} studentized deleted residual.
- It can be shown that $t_i = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$.

  Further since $\underbrace{(n-p)MSE}_{SSE} = (n-p-1)MSE_{(i)} + \frac{e_i^2}{1-h_{ii}}$,

$\Sigma(y_i - \hat{y}_i)^2$

$$\Rightarrow t_i = e_i\left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}\right]^{1/2}$$ — using full data set.

Test for Outliers:

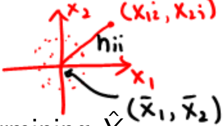
- First identify the cases with large $|t_i|$.
- Then use the following rule:
  - $|t_i| \leq t(1-\alpha/2n : n-p-1)$ $\Rightarrow$ case i is not an outlier.
  - $|t_i| > t(1-\alpha/2n : n-p-1)$ $\Rightarrow$ case i is an outlier.

RR

where $t(1-\alpha/2n : n-p-1)$ is called the Bonferroni critical value.

Identifying Outliers with respect to X

Hat matrix (the diagonal elements h_{ii} of H (leverage values)) can be used to identify outliers with respect to X. Here h_{ii} is the distance between the value of X for i^{th} observation and the center of X.



Properties of leverage values (h_{ii}) :

- $0 \leq h_{ii} \leq 1$, $\quad \sum_{i=1}^{n} h_{ii} = p$.
- Since $\hat{Y} = HY$, h_{ii} : the weight of Y_i in determining $\hat{Y}_i$.
- The larger is h_{ii}, the smaller is $\sigma^2\{e_i\}$ $= (1-h_{ii})\sigma^2$ $(h_{ii} = 1 \Rightarrow \sigma^2\{e_i\} = 0)$.

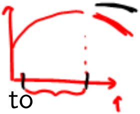
Test for Outliers:

- First identify the cases with large leverage values.
- Then use the following rule:
  - $h_{ii} \leq 2\bar{h} = 2p/n \;\Rightarrow$ case i is not an outlier,
  - $h_{ii} > 2\bar{h} = 2p/n \;\Rightarrow$ case i is an outlier,

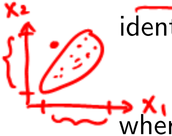
where $\bar{h} = \frac{\sum h_{ii}}{n} = \frac{p}{n}$ and $\frac{2p}{n} \leq 1$. .

Identifying Hidden Extrapolation

- When there are only two predictor variables, a scatter plot can be used to identify extrapolation. This simple graphic analysis is no longer available with larger numbers of predictor variables, where extrapolations may be hidden.



- Leverage values for new set of X_{new} values can be used to identify hidden extrapolations.



$$h_{new,new} = X'_{new}(X'X)^{-1}X_{new},$$

where X_{new} is the vector containing the new X observation.

Rule:

If $h_{new,new}$ is in the range of h_{ii}s for the cases in the data set, no extrapolation is involved. On the other hand, if $h_{new,new}$ is much larger than the leverage values for the cases in the data set, an extrapolation is indicated.

Identifying Influential Cases

- A case is influential if its exclusion causes major changes in the fitted regression function.
- Not all outlying cases are influential.
- We take up three measures of influence that are widely used in practice.
  1. Influence on Single Fitted Value-DFFITS
  2. Influence on All Fitted Values-Cook's Distance
  3. Influence on the Regression Coefficients-DFBETAS

.

Influence on Single Fitted Value-*DFFITS*

- DFFITS is the difference between the fitted value $\hat{Y}_i$ (with all n cases) and $\hat{Y}_{i(i)}$ (without the i^{th} case) in terms of standard deviations.

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

- It can be shown that the *DFFITS* values can be computed by using only the results from fitting the entire data set, as follows:

$$(DFFITS)_i = t_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2} \text{— full data set.}$$

Rule: The i^{th} case is influential if

- $|DFFITS| > 1$ for for small to medium data sets, $(n \leq 36)$
- $|DFFITS| > 2\sqrt{p/n}$ for large data sets. $(n > 36)$

Influence on All Fitted Values-*Cook's Distance*

- Cooks distance (D_i) considers the influence of the ith case on all n fitted values.

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{pMSE} = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(\mathbf{i})})(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(\mathbf{i})})}{pMSE}.$$

- It can be shown that

$$D_i = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$ — using the full data set.

- D_i depends on two factors: (1) the size of the residual ei and (2) the leverage value h_{ii} ($e_i \uparrow$ or $h_{ii} \uparrow \Rightarrow D_i \uparrow$).

Rule: $D_i \sim F(p, n-p)$

First identify the cases with large Cook's Distances.

- little influence if $P(D_i \leq d_i^*) > 0.2$,
- major influence if $P(D_i \leq d_i^*) > 0.5$,

where d_i^* is a observed value of D_i.

Influence on the Regression Coefficients-*DFBETAS*

- *DFBETAS* considers the influence on regression coefficients: the difference between b_k (with all n cases) and $b_{k(i)}$ (without the i^{th} case) in terms of standard deviations.

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}c_{kk}}}, \quad k = 0, 1, \cdots, p-1,$$

where c_{kk} : the k^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.

Rule: The i^{th} case is influential if

- $|DFBETAS| > 1$ for for small to medium data sets,
- $|DFBETAS| > 2\sqrt{n}$ for large data sets.

.

Multicollinearity

There are some key problems that typically arise when the predictor variables of the regression model are highly correlated among themselves:

1. Adding or deleting a predictor variable changes the regression coefficients,
2. The extra sum of squares associated with a predictor variable varies, depending upon which other predictor variables are already included in the model. $ESS(X_1|X_2) \approx ESS(X_1|X_3)$
3. The estimated standard deviations of the regression coefficients become large $S\{b_k\}$
   makes the C.I wider
4. The estimated regression coefficients individually may not be statistically significant even though a definite statistical relation exists between the response variable and the set of predictor variables. $Y = \beta_1 X_1 + \beta_2 X_2$
   >0
   <0

Multicollinearity Informal Diagnostics

1. Large changes in the estimated regression coefficients when a predictor variable is added or deleted, or when an observation is altered or deleted.
2. Nonsignificant results in individual tests on the regression coefficients for important predictor variables.
3. Estimated regression coefficients with an algebraic sign that is the opposite of that expected from theoretical considerations or prior experience.
4. Large coefficients of simple correlation between pairs of predictor variables in the correlation matrix r_{xx}.
5. Wide confidence intervals for the regression coefficients representing important predictor variables.

Limitations: do not provide quantitative measurements, may not identify the nature of the multicollinearity

Variance Inflation Factor (VIF)

Use of variance inflation factors is widely accepted and formal multicollinearity diagnostics method. $(VIF)_k$ tells us which predictors are highly correlated with other predictors.

$$(VIF)_k = (1 - R_k^2)^{-1}, \quad k = 1, 2, \ldots, p-1,$$

where where R_k^2 is the coefficient of multiple determination when X_k is regressed on the $p-2$ other X variables in the model.

response $(X_k \sim X_1, X_2, \ldots X_{k-1}, X_{k+1} \ldots, X_{p-1})$

Rule:

If one or more $VIF_k > 10$, we may want to eliminate some of the predictors.

.