Linear Statistical Modeling Methods with SAS

Model Diagnostics

Xuemao Zhang
East Stroudsburg University

March 15, 2024

Outline

- Residual Analysis to Check Model Assumptions
- Added Variable Plots
- Detecting Outliers and Influential Points
  - Studentized Residuals
  - PRESS Residuals
  - Studentized Deleted Residuals
  - Hat Matrix
  - DFFITS
  - Cook's Distance
  - DFBETAS
- Detecting Multicollinearity

Residual Analysis to Check Model Assumptions

residual

$$e_i = y_i - \widehat{y}_i, \quad i = 1, \ldots, n.$$

Note. $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \widehat{y}_i) = 0.$

Recall that in matrix notation, $\mathbf{e} = (e_1, \ldots, e_n)'$.

$$\boxed{\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}, \quad \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.}$$

$$\boxed{Var(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H}).}$$

That is,

$$\boxed{Var(e_i) = \sigma^2(1 - h_{ii}), \quad Cov(e_i, e_j) = -h_{ij}\sigma^2}$$

Note that residuals are **not** independent.

A residual plot is a graph of the points

$$(\widehat{y}_i, y_i - \widehat{y}_i), \quad i = 1, \ldots, n.$$

or

$$(x_{ji}, y_i - \widehat{y}_i), \quad i = 1, \ldots, n, j = 1, \ldots, k,$$

which is the plot of residuals versus each regressor $X_j, j = 1, \ldots, k$.

Residual Analysis to Check Model Assumptions

When analyzing a residual plot, look for a pattern in the way the points are configured, and use these criteria:

(1) **Independence:** The residual plot should not have any obvious patterns (not even a straight line pattern). This confirms that the scatterplot of the sample data is a straight-line pattern.

(2) **Constance Variance:** The residual plot should not become thicker (or thinner) when viewed from left to right. This confirms the requirement that for different fixed values of X variables, the distributions of the corresponding Y values all have the same standard deviation.

(3) **Normality.** Residuals should be plotted on a normal quantile plot (the SAS procedure univariate will do this). The population distribution is Normal if the pattern of the points is reasonably close to a straight line and the points do not show some **systematic** pattern that is not a straight-line pattern.

- Please check Lecture10_MLR_Part1.pdf to see the SAS code for residual analysis.

Added Variable Plots

- Building a regression model is **variable selections**. That is, what predictors should be included in the regression model.
- The purpose of the added variable plot is illustrate the **effect of adding each regressor** to a model that contains all the others.
- Consider a full data set with k regressors

$$(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i), i = 1, \ldots, n.$$

Suppose we want to check the effect of the regressor X_1.

Step 1. Fit the regresson model using $X_2, \ldots, X_k$ only:

$$\widehat{E(Y)} = \widehat{\beta}_0 + \widehat{\beta}_2 X_2 + \cdots + \widehat{\beta}_k X_k,$$

and get the residuals

$$e_i(Y|X_2, X_3, \ldots, X_k) = y_i - \widehat{y}_i(X_2, X_3, \ldots, X_k), i = 1, \ldots, n.$$

Step 2. We also regard X_1 as a response and regress X_1 on $X_2, X_3, \ldots, X_k$. Suppose the fitted model is

$$\widehat{E(X_1)} = \widehat{\beta^*}_0 + \widehat{\beta^*}_2 X_2 + \cdots + \widehat{\beta^*}_k X_k,$$

and get the residuals

$$e_i(X_1|X_2, X_3, \ldots, X_k) = x_{1i} - \widehat{x}_{1i}(X_2, X_3, \ldots, X_k), i = 1, \ldots, n.$$

Added Variable Plots

Added-variable plots

The added-variable plot for the regressor variable X_1 consists of a plot of the Y residuals $e(Y|X_2, X_3, \ldots, X_k)$ against the X_1 residuals $e(X_1|X_2, X_3, \ldots, X_k)$.

- Linear trend of the plot indicates that the model should include a linear term in X_1.
- Curvilinear pattern of the plot indicates that the model should include a more complex function of X_1.
- No pattern (or a horizontal one) indicate that X_1 should not be included.
- We may also can identify some extreme data points or non-constant variance based on these plots.

Example

- Table 10.1 shows a portion of the data on average annual income of managers during the past two years (X_1), a score measuring each manager's risk aversion (X_2), and the amount of life insurance carried (Y) for a sample of 18 managers in the 30-39 age group. Risk aversion was measured by a standard questionnaire administered to each manager: the higher the score, the greater the degree of risk aversion. Income and risk aversion are mildly correlated here, the coefficient of correlation being $r_{12} = 0.254$.

- Next lecture slide shows the data.

Example

```
data ch10tab01;
input x1 x2 y@@;
label x1 = 'Income'
x2 = 'Risk Aversion'
y = 'Insurance';
cards;
45.010 6 91 57.204 4 162
26.852 5 11 66.290 7 240
40.964 5 73 72.996 10 311
79.380 1 316 52.766 8 154
55.916 6 164 38.122 4 54
35.840 6 53 75.796 9 326
37.408 5 55 54.376 2 130
46.186 7 112 46.130 4 91
30.366 3 14 39.060 5 63
;
run;
```

Example

```
proc reg data = ch10tab01 ;
model y = x1 x2 / partial ; /*partial regression plots*/
plot r.*x1 r.*x2; /* Partial regression plots for x1 and x2 */
/*residuals vs x1 and residuals vs x2*/
run;
```

- **Added variable plots** is also called **partial regression plot**.
- The partial residual plot clearly suggests that a linear relation for X_1, is not appropriate in the model already containing X_2.

Example

- Generate Added variable plots or partial regression plots manually.

```
proc reg data=ch10tab01 noprint; /* fit without output*/
model y x1 = x2 ;
output out=tempx1 r=ry rx;
run;

proc gplot data=tempx1;
plot ry*rx;
label ry='e(Y|X2)'
rx='e(X1|X2)';
run;
```

- Or we use `proc sgplot`.

```
proc sgplot data=tempx1;
scatter x=rx y=ry;
label ry='e(Y|X2)'
rx='e(X1|X2)';
run;
```

Detecting Outliers and Influential Points

Frequently in regression analysis applications, the data set contains some cases that are out-lying or extreme. There are three cases

- Outliers only: Extreme in X but its Y consistent with (not extreme in Y) the regression model.
- Influential Points: Extreme in Y. It is strongly inconsistent with the regression model. The points are outliers as well (may or may not be extreme in X).

- We use several types of residuals to detect influential points

Studentized Residuals

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

$$s\{e_i\} = \widehat{\sigma}(e_i) = \sqrt{MSE(1 - h_{ii})}$$

studentized residual

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

While $Var(e_i) = \sigma^2(1 - h_{ii})$, the studentized residuals, standardized residuals, r_i have constant variance. Studentized residuals often are called **internally studentized residuals**.

PRESS Residuals

PRESS residuals are also called **Deleted Residuals**. **PRESS residuals** are computed as follows:

- First, observation (X_i, Y_i) is omitted from the data and the least squares line fit to the remaining data, giving the parameter estimates $\hat{\beta}_0^{(i)}, \hat{\beta}_1^{(i)}, \ldots, \hat{\beta}_k^{(i)}$.
- Next, the **deleted fitted value**, $\hat{Y}_{(i)} = \hat{\beta}_0^{(i)} + \hat{\beta}_1^{(i)} X_1 i + \cdots \hat{\beta}_k^{(i)} X_{ki}$ is computed, $i = 1, \ldots, n$.
- Then, the **deleted residual** or **PRESS residuals** $e_{(i)} = Y_i - \hat{Y}_{(i)}$ is computed, $i = 1, \ldots, n$.

In fact, the PRESS residual can be calculated from the ordinary residual. Namely,

$$e_{(i)} = \frac{y_i - \widehat{y}_i}{1 - \mathbf{x_i}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}} = \frac{e_i}{1 - h_{ii}}.$$

PRESS Residuals

Note that a deleted residual also corresponds to the prediction error for a new observation. So an estimate standard deviation of $e_{(i)}$ is

$$s\{e_{(i)}\} = \widehat{\sigma}_{e_{(i)}} = \sqrt{MSE_{(i)}(1 + \mathbf{X_i}'(\mathbf{X}'_{(\mathbf{i})}\mathbf{X}_{(\mathbf{i})})^{-1}\mathbf{X_i})},$$

where $MSE_{(i)}$ is the mean square error when the ith case is omitted in fitting the regression function, and $X_{(i)}$ is the X matrix with the ith case deleted. It can be shown that

$$s^2\{e_{(i)}\} = \widehat{\sigma}_{e_{(i)}} = \frac{MSE_{(i)}}{1 - h_{ii}}$$

Furthermore,

$$\boxed{\frac{e_{(i)}}{s\{e_{(i)}\}} \sim t(n - 2 - k).}$$

Thus, the standardized PRESS residuals should be plotted on quantile plot for the t_{n-2-k} distribution to check the model Normality assumption.

Studentized Deleted Residuals

Combining the studentized residuals and PRESS residuals, we call

$$t_i = \frac{e_{(i)}}{s\{e_{(i)}\}}$$

the **studentized deleted residual**. Or an equivalent expression

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

The studentized deleted residual t_i is also called an **externally studentized residual**, in contrast to the **internally studentized residual** r_i.

Again, t_i can be calculated without having to fit new regression functions each time a different case is omitted.

$$t_i = e_i \left[\frac{n - 2 - k}{SSE(1 - h_{ii}) - e_i^2} \right]^{1/2}$$

Use of Hat Matrix

The hat matrix, as we saw, plays an important role in determining the magnitude of a studentized deleted residual and therefore in identifying outlying Y observations. The hat matrix also is helpful in directly identifying outlying X observations. The diagonal elements h_{ii} of the hat matrix have some useful properties, especially

$$\boxed{0 \leq h_{ii} \leq 1, \quad \sum_{i=1}^{n} h_{ii} = k + 1.}$$

- h_{ii} is the weight of observation Y_i in determining this fitted value. Therefore, h_{ii} measures the role of the X values in determining how important Y_i is in affecting the fitted value $\widehat{Y}_i$.
- Hence, the larger is h_{ii}, the closer the fitted value $\widehat{Y}_i$ will tend to be to the observed value Y_i.
- $Var(e_i) = \sigma^2(1 - h_{ii})$; $Var(e_i) = \sigma^2(1 - h_{ii})$.

Remark. A leverage value h_{ii} is usually considered to be large if it is more than twice as large as the mean leverage value $\frac{k+1}{n}$. Hence, leverage values greater than $\frac{k+1}{n}$ are considered by this rule to indicate outlying cases with regard to their X values

DFFITS - Influence on the fitted value

Recall that the PRESS residual is

$$e_{(i)} = Y_i - \hat{Y}_{(i)} = \frac{e_i}{1 - h_{ii}}, i = 1, \ldots, n.$$

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

It is computed as

$$(\text{DFFITS})_i = e_i \left[\frac{n - 2 - k}{SSE(1 - h_{ii}) - e_i^2)} \right]^{1/2} \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2} = t_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2},$$

where t_i is the studentized deleted residual.

A larger h_{ii} makes this expression large. Therefore, a

$$|(\text{DFFITS})_i| \geq 1$$

is considered extreme for small to mediate data sets and $|(\text{DFFITS})_i| \geq 2\sqrt{(k+1)/n}$ is considered extreme for large data sets.

Cook's Distance - Influence on all fitted values

In contrast to the DFFITS measure which considers the influence of the ith case on the fitted value Y_i for this case, Cook's distance measure considers the influence of the ith case on all n fitted values.

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)} \right)^2}{(k+1)MSE}$$

The Cook's distance can be expressed as

$$D_i = \frac{\left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(\mathbf{i})} \right)' \left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(\mathbf{i})} \right)}{(k+1)MSE},$$

where Here, $\hat{\mathbf{Y}}$ is the vector of the fitted values when all n cases are used for the regression fit and $\hat{\mathbf{Y}}_{(i)}$ is the vector of the fitted values when the ith case is deleted.

Cook's Distance - Influence on all fitted values

Cook's distance measure D_i can be calculated without fitting a new regression function each time a different case is deleted. An algebraically equivalent expression is

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

A D_i is considered large if

$$D_i \geq F_{0.25,k+1,n-1-k} \quad \text{(upper quartile of the F distribution).}$$

This criterion identifies a data point as influential if its $|e_i|$ is large and/or its h_{ii} is large. Some researchers use the conventional cut-off point $4/n$.

DFBETAS - Influence on the Regression Coefficients

$$\text{DEBETAS}_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{MSE_{(i)} c_{jj}}},\ j = 0, 1, \ldots, k,$$

where $\hat{\beta}_{j(i)}$ is the jth regression coefficient computed without the use of the ith observation and c_{jj} is the jth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.

Recall the the variance of $\hat{\beta}_j$ is

$$Var\{\hat{\beta}_j\} = c_{jj}\sigma^2$$

The error term variance σ^2 here is estimated by $MSE_{(i)}$, the error mean square obtained when the ith case is deleted in fitting the regression model.

A DFBETA displays the (studentized) change in the jth estimated slope when the ith data point is deleted from the data. It is extreme if

$$|\text{DEBETAS}_{j(i)}| \geq 1$$

for small/medium sized data sets and $|\text{DEBETAS}_{j(i)}| \geq 2/\sqrt{n}$ is considered extreme for large data sets.

Example

```
data ch9tab01;
input x1 x2 x3 x4 y@@;
logy = log(y);
label x1 = 'blood-clotting'
x2 = 'prognostic'
x3 = 'enzyme'
x4 = 'liver function'
y = 'survival';
cards;
6.7  62   81  2.59  200 5.1  59   66  1.70  101
7.4  57   83  2.16  204 6.5  73   41  2.01  101
7.8  65  115  4.30  509 5.8  38   72  1.42   80
5.7  46   63  1.91   80 3.7  68   81  2.57  127
6.0  67   93  2.50  202 3.7  76   94  2.40  203
6.3  84   83  4.13  329 6.7  51   43  1.86   65
5.8  96  114  3.95  830 5.8  83   88  3.95  330
7.7  62   67  3.40  168 7.4  74   68  2.40  217
6.0  85   28  2.98   87 3.7  51   41  1.55   34
7.3  68   74  3.56  215 5.6  57   87  3.02  172
5.2  52   76  2.85  109 3.4  83   53  1.12  136
6.7  26   68  2.10   70 5.8  67   86  3.40  220
6.3  59  100  2.95  276 5.8  61   73  3.50  144
5.2  52   86  2.45  181 1.2  76   90  5.59  574
5.2  54   56  2.71   72 5.8  76   59  2.58  178
3.2  64   65  0.74   71 8.7  45   23  2.52   58
5.0  59   73  3.50  116 5.8  72   93  3.30  295
5.4  58   70  2.64  115 5.3  51   99  2.60  184
2.6  74   86  2.05  118 4.3   8  119  2.85  120
4.8  61   76  2.45  151 5.4  52   88  1.81  148
5.2  49   72  1.84   95 3.6  28   99  1.30   75
8.8  86   88  6.40  483 6.5  56   77  2.85  153
3.4  77   93  1.48  191 6.5  40   84  3.00  123
4.5  73  106  3.05  311 4.8  86  101  4.10  398
5.1  67   77  2.86  158 3.9  82  103  4.55  310
6.6  77   46  1.95  124 6.4  85   40  1.21  125
6.4  59   85  2.33  198 8.8  78   72  3.20  313
;
run;
```

Example

- SAS has an influence option in the model statement that will display most of the diagnostic statistics.

```
proc reg data = ch9tab01;
model logy= x1 - x4/ influence;
output out=RegOut cookd=CooksD; /*dffits=DFFits*/
run;
```

```
proc print data=RegOut;
run;
```

Detecting Multicollinearity - VIF

Recall that when multicollinearity is present, we consider the **correlation transformation**

correlation transformation

$$Y_j^* = \frac{1}{\sqrt{n-1}} \frac{Y_j - \overline{Y}}{S_Y}, j = 1, \ldots, n$$

$$x_{ij}^* = \frac{1}{\sqrt{n-1}} \frac{x_{ij} - \overline{X_i}}{S_{X_i}}, j = 1, \ldots, n, i = 1, \ldots, k.$$

Then $\mathbf{X}^{*'}\mathbf{X}^*$ is a $k \times k$ correlation matrix

$$\mathbf{X}^{*'}\mathbf{X}^* = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1,k-1} & r_{1k} \\ r_{21} & 1 & \cdots & r_{2,k-1} & r_{2k} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{k-1,1} & r_{k-1,2} & \cdots & 1 & r_{k-1,k} \\ r_{k1} & r_{k2} & \cdots & r_{k,k-1} & 1 \end{bmatrix}$$

Detecting Multicollinearity - VIF

Thus the least squares estimator of β^*

$$\widehat{\beta^*} = (\mathbf{X}^{*'}\mathbf{X}^*)^{-1}\mathbf{X}^{*'}\mathbf{Y}^*$$

is reduced to

$$\widehat{\beta^*} = \mathbf{r_{XX}}^{-1}\mathbf{r_{YX}},$$

where $\mathbf{r_{XX}} = \mathbf{X}^{*'}\mathbf{X}^*$ and $\mathbf{r_{YX}}$ is a vector containing the coefficients of simple correlation between the response variable Y and each of the X variables, denoted by r_{Y1}, r_{Y2}, etc.:

$$\mathbf{r_{YX}} = \begin{bmatrix} r_{Y1} \\ r_{Y2} \\ \vdots \\ r_{Yk} \end{bmatrix}$$

Hence,

$$Var(\widehat{\beta^*}) = (\sigma^*)^2\mathbf{r_{XX}}^{-1},$$

where $(\sigma^*)^2$ is the error term variance for the transformed model.

Detecting Multicollinearity - VIF

Therefore,

$$Var(\widehat{\beta}_j^*) = (\sigma^*)^2 (VIF)_j$$

where $(VIF)_j$ denotes the jth diagonal element of the matrix $\mathbf{r_{XX}}^{-1}$.

$(VIF)_j$ is called the **variance inflation factor** (VIF) for $\widehat{\beta}_j^*$.

It can be shown that

$$(VIF)_j = (1 - R_j^2)^{-1},\ j = 1, 2, \ldots, k$$

where R_j^2 is the coefficient of multiple determination when X_j is regressed on the $k - 1$ other X variables in the model.

The largest VIF value among all X variables is often used as an indicator of the severity of multicollinearity. It is generally believed that if any VIF exceeds 10, there is a reason for at least some concern.

Example

```
proc reg data = ch9tab01;
model logy = x1 - x4/ vif tol stb collin;
run;
```

- `VIF`: Computes Variance Inflation Factor (VIF) to assess multicollinearity.
- `TOL`: Computes Tolerance statistics to assess multicollinearity.
- `STB`: Requests Standardized Beta Coefficients in the output.
- `COLLIN`: Requests Collinearity diagnostics in the output.

Example

- Check the correlation matrix

```
proc corr data=ch9tab01;
var x1 x2 x3 x4;
run;
```

License

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.