Chapter 9

Building the Regression Model II: Diagnostics

9.1 Model Adequacy For A Predictor Variable– Partial Regression Plots

SAS program: att8-9-1-read-partial-res-plot

Partial regression plots (added variable plots, adjusted variable plots) are an aid to identify the nature and strength of the marginal relation for a predictor X_i, given the other predictors are already in the model. Partial residual plots are also useful in identifying outliers.

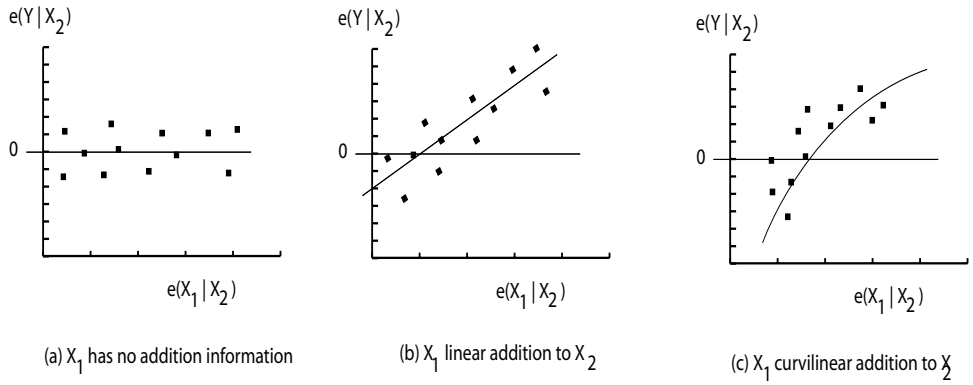


Figure 9.1 (Example Partial Regression Plots)

Figure (a) above indicates predictor variable X_1 provides no additional information over and above X_2, to predict Y. Figures (b) and (c), on the other hand, indicate that X_1 provide an additional linear and curvilinear effect, respectively, over and above X_2, to predict Y.

Exercise 9.1 (Model Adequacy For A Predictor Variable–Partial Regression Plots)

illumination, X_{i1}	9	7	11	16	21	19	23	29	31	33
noise, X_{i2}	15	20	17	22	24	26	28	30	29	37
ability to read, Y	70	70	75	88	91	94	100	92	90	85

Here, the first–order regression model gives

$$\hat{Y} = 69.5 + 0.75X_1 + 0.04X_2$$

1. *(Partial) Regression, Y on X_1.*
   The regression of Y on X_1 (alone) is (circle one)
   $\boldsymbol{\hat{Y}(X_2) = 70.01 + 0.78X_1}$
   $\boldsymbol{\hat{Y}(X_2) = 70.01 + 0.78X_2}$
   $\boldsymbol{\hat{Y}(X_2) = 60.30 + 0.75X_2}$

2. *Partial Residuals, $e(Y|X_1)$.*
   True / False
   The residuals for the regression of Y on X_1, $e(Y|X_1)$, are

illumination, X_{i1}	9	7	11	16	21	19	23	29	31	33
ability to read, Y	70	70	75	88	91	94	100	92	90	85
residuals, $Y - \hat{Y}(X_1)$	-7.02	-5.46	-3.57	5.54	4.64	9.20	12.09	-0.58	-4.14	-10.70

3. *(Partial) Regression, Y on X_2.*
   The regression of Y on X_2 (alone) is (circle one)
   $\boldsymbol{\hat{Y}(X_2) = 69.5 + 0.75X_1}$
   $\boldsymbol{\hat{Y}(X_2) = 60.30 + 1.02X_2}$
   $\boldsymbol{\hat{Y}(X_2) = 60.30 + 0.75X_2}$

4. *Partial Residuals, $e(Y|X_2)$.*
   True / False
   The residuals for the regression of Y on X_2, $e(Y|X_2)$, are

noise, X_{i2}	15	20	17	22	24	26	28	30	29	37
ability to read, Y	70	70	75	88	91	94	100	92	90	85
residuals, $Y - \hat{Y}(X_2)$	-5.54	-10.62	-2.57	5.35	6.31	7.28	11.25	1.22	0.23	-12.90

5. *(Partial) Regression, X_1 on X_2.*
   The regression of X_1 on X_2 (alone) is (circle one)
   $\boldsymbol{\hat{X}_1(X_2) = 69.5 + 0.75X_1}$
   $\boldsymbol{\hat{X}_1(X_2) = 60.30 + 1.02X_2}$
   $\boldsymbol{\hat{X}_1(X_2) = -12.29 + 1.30X_2}$

6. *Partial Residuals, $e(X_1|X_2)$.*
   True / False
   The residuals for the regression of X_1 on X_2, $e(X_1|X_2)$, are

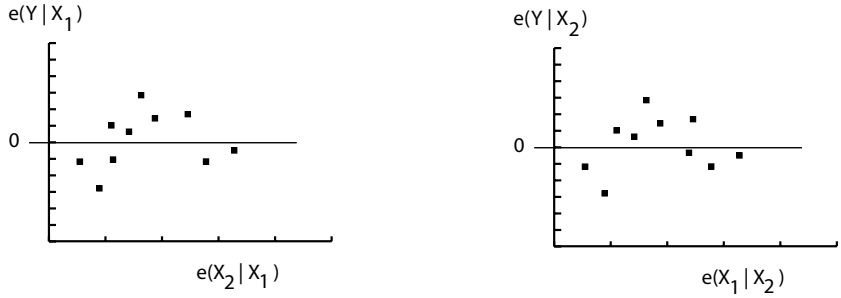
illumination, X_{i1}	9	7	11	16	21	19	23	29	31	33
noise, X_{i2}	15	20	17	22	24	26	28	30	29	37
residuals, $X_1 - \hat{X}_1(X_2)$	1.82	-6.67	1.22	-0.27	2.14	-2.46	-1.05	2.35	5.65	-2.73

7. *(Partial) Regression, X_2 on X_1.*
The regression of X_2 on X_1 (alone) is (circle one)
$\boldsymbol{\hat{X}_2(X_1) = 69.5 + 0.75X_1}$
$\boldsymbol{\hat{X}_2(X_1) = 11.58 + 0.66X_1}$
$\boldsymbol{\hat{X}_2(X_1) = -12.29 + 1.30X_2}$

8. *Partial Residuals, $e(X_2|X_1)$.*
True / **False**
The residuals for the regression of X_2 on X_1, $e(X_2|X_1)$, are

illumination, X_{i1}	9	7	11	16	21	19	23	29	31	33
noise, X_{i2}	15	20	17	22	24	26	28	30	29	37
residuals, $X_1 - \hat{X}_1(X_2)$	-2.56	3.77	-1.89	-0.21	-1.53	1.80	1.14	-0.85	-3.17	3.50

9. *Partial Residual Plots*



(a) partial regression plot to determine X_2 influence

(b) partial regression plot to determine X_1 influence

Figure 9.2 (Partial Regression Plots)

There are two partial residual plots (choose two)

(a) $e(Y|X_1)$ versus $e(X_2|X_1)$

(b) $e(Y|X_2)$ versus $e(X_1|X_2)$

(c) $e(X_1|X_2)$ versus $e(X_2|X_1)$

The residual plot in (a), $e(Y|X_1)$ versus $e(X_2|X_1)$, indicates that
(circle one) **no** / **a linear** / **a curved**
X_2 variable should be added to a model already containing variable X_1.

The residual plot in (b), $e(Y|X_2)$ versus $e(X_1|X_2)$, indicates that
(circle one) **no** / **a linear** / **a curved**
X_1 variable should be added to a model already containing variable X_2.

In other words, once one (either one) of the two variables are in the model, the other variable need not be added.

9.2 Identifying Outlying Y Observations–Studentized Deleted Residuals

SAS program: att8-9-2-read-student-del-resid

We describe how to identify *outlying* (not influential) Y (not X) observations in this section. A scatter plot, where some of the data points are either outlying Y observations or outlying X observations, is given below.

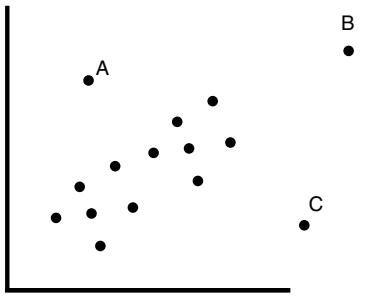


Figure 9.3 (Outlying Y and X Observations)

Observation A is a outlying (circle one or two) ***X*** / ***Y*** observation.
Observation B is a outlying (circle one or two) ***X*** / ***Y*** observation.
Observation C is a outlying (circle one or two) ***X*** / ***Y*** observation.
Of the three points, only point (choose one) **A** / **B** / **C** is *NOT influential* in the sense although it is an outlier, the regression function was "headed its way" in any case.

Outlying influential Y observations can be identified with large (in absolute value) studentized deleted residuals. Studentized deleted residuals, t_i, are calculated in the following way,

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

where *MSE* is the mean sums of squared error (from the ANOVA associated with the regression), h_{ii} is a diagonal element of the leverage (hat) matrix and e_i is the ith residual.

Exercise 9.2 (Identifying Outlying Y Observations–Studentized Deleted Residuals)

observation, i	1	2	3	4	5	6	7	8	9	10
illumination, X_{i1}	9	7	11	16	21	19	23	29	31	33
noise, X_{i2}	15	20	17	22	24	26	28	30	29	37
ability to read, Y	70	70	75	88	91	94	100	92	90	85

1. *Largest Studentized Deleted Residual.*
   The largest studentized deleted residual (RStudent in SAS output) is observation (choose one) **5** / **7** / **10**
   where
   $$t_{10} = -2.5242$$

   The studentized deleted residual associated with observation 10, appears to be a Y outlier because it is over two and a half (studentized deleted residual) standard deviations (2.5242 to be exact) from zero.

2. *Bonferroni outlier test procedure.*
   Test if observation 10 is an outlier or not in the Y direction at $\alpha = 0.10$.

   (a) *Statement.*
   The statement of the test is (check none, one or more):

   i. $H_0 : \beta_1 = \beta_2 = 0$ versus $H_a :$ $\beta_1 > 0, \beta_2 \neq 0$.
   ii. $H_0 : \beta_1 = \beta_2 = 0$ versus H_a : not all β_i equal to zero.
   iii. H_0 : observation Y_{10} is *not* an outlier versus H_a : it is an outlier

   (b) *Test.*
   The test statistic is $|t_{10}| = |-2.52| = 2.52$
   The Bonferroni critical value at $\alpha = 0.10$ is
   $t(1 - \alpha/2n; n - p - 1) = t(1 - 0.10/2(10); 10 - 3 - 1) =$
   (choose one) **3.71** / **4.32** / **5.99**
   (Use PRGM INVT ENTER 6 ENTER 0.995 ENTER)

   (c) *Conclusion.*
   Since the test statistic, 2.52, is smaller than the critical value, 3.71, we (circle one) **accept** / **reject** the null hypothesis that observation 10 is *not* an outlier in the Y direction.

3. **True** / **False**
   Since observation 10, with the largest studentized deleted residual value, is not an outlier, then all of the other observations are also not outliers in the Y direction.

4. *Understanding the Studentized Deleted Residuals*
   The error, e_i, is given by
   $$e_i = Y_i - \hat{Y}_i$$

and so an observation with large error is one which is a (choose one) **large** / **small** vertical (Y) distance from the regression line.

The (internally) *studentized residual* is given by

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

and so the size of an error, e_i, is measured relative to the standard deviation in the error, $s\{e_i\}$. An observation where r_i is larger than (choose one) **one** / **two** / **seven** is usually considered an outlier.

The *studentized deleted residual* is given by

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

is the identical to r_i except $MSE_{(i)}$ in this statistic is based on all the observations *except* the ith one. Deleting the ith observation while calculating $MSE_{(i)}$ and measuring e_i relative to a function of this quantity tends to highlight the influence of the ith observation.

9.3 Identifying Outlying X Observations–Hat Matrix Leverage Values

SAS program: att8-9-3-read-leverage

Outlying X observations (not Y observations!) are identified with large (in absolute value) diagonal elements of the leverage (hat) matrix, h_{ii}, called *leverage* values.

Exercise 9.3 (Identifying Outlying X Observations–Hat Matrix Leverage Values)

i	1	2	3	4	5	6	7	8	9	10
illumination, X_{i1}	9	7	11	16	21	19	23	29	31	33
noise, X_{i2}	15	20	17	22	24	26	28	30	29	37
ability to read, Y	70	70	75	88	91	94	100	92	90	85

1. *Observations With Largest Leverages.*
   The observation with the largest leverage (Hat Diag H) is
   observation (choose one) **2** / **9** / **10**

2. *Outlying Influential X Observations.*
   A rule of thumb is if
   $$h_{ii} > \frac{2p}{n}$$
   then the corresponding X values are outliers. Since $h_{ii} > 2p/n = 2(3)/10 = 0.6$, the following observations are outliers:
   (circle none, one or more) **none** / **2** / **10**.

3. *Assessing Whether New Observations Are X Outliers.*
   New observations are outliers if
   $$h_{new,new} = \mathbf{X}'_{\mathbf{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_{\mathbf{new}}$$
   is not within range of the other h_{ii}. For example, if
   $$\mathbf{X}'_{new} = [1, 23, 39]$$
   then from SAS, $h_{new,new}$ = (circle one) **1.53** / **2.14** / **2.83**
   which is much larger than all of the other h_{ii} values,
   and so $\mathbf{X}' = [1, 23, 39]$ is an outlier.

4. **True** / **False**
   $\sum h_{ii} = p = 3$, the number of parameters in the regression equation.

5. **True** / **False**
   $0 \leq h_{ii} \leq 1$.

9.4 Identifying Influential Cases–*DFFITS*, Cook's Distance and *DFBETAS* Measures

SAS program: att8-9-4-read-influential

An X or Y observation is *influential* (as opposed to simply outlying) if its exclusion causes major changes to the regression function. Three measures of influence are considered here (all are based on omitting the i case and determining its influence): *DFFITS*, Cook's Distance and *DFBETAS* Measures.

Exercise 9.4 (Identifying Influential Cases–*DFFITS*, Cook's Distance and *DFBETAS* Measures)

i	1	2	3	4	5	6	7	8	9	10
illumination, X_{i1}	9	7	11	16	21	19	23	29	31	33
noise, X_{i2}	15	20	17	22	24	26	28	30	29	37
ability to read, Y	70	70	75	88	91	94	100	92	90	85

i	$(DFFITS)_i$	D_i	$(DFBETAS)_{k(i)}, b_0$	$(DFBETAS)_{k(i)}, b_1$	$(DFBETAS)_{k(i)}, b_2$
1	-0.7990	0.211	-0.6892	-0.2304	0.4532
2	-1.1964	0.474	0.3319	1.0193	-0.8051
3	-0.2737	0.028	-0.2203	-0.0627	0.1353
4	0.2470	0.022	0.0994	-0.0184	-0.0202
5	0.2342	0.020	0.1395	0.1278	-0.1279
6	0.5285	0.087	-0.1955	-0.3147	0.3219
7	0.6853	0.122	-0.2362	-0.1897	0.2872
8	-0.0359	0.000	-0.0019	-0.0174	0.0088
9	-0.5402	0.107	-0.2257	-0.4439	0.3486
10	-2.7827	1.460	2.1073	0.9972	-1.7837

1. *Influence Single i Has On Single $\hat{Y}_i$:* $(DFFITS)_i$.

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

If this measure is *large* (bigger than 1 for small to medium data sets or bigger than $2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{3}{10}} \approx 1.1$ for large data sets), then the corresponding observation is influential. For the reading versus illumination data, the influential points are (look at the $(DFFITS)_i$ values above)
(*choose two!*) **2** / **8** / **10**

2. *Influence Single i Has On All $\hat{Y}$: Cook's Distance D_i.*

$$D_i = \frac{e_i^2}{pMSE} \frac{h_{ii}}{(1 - h_{ii})^2}$$

If this measure is *large* (the *percentile* associated with this measure in a $F(p, n - p) = F(3, 10 - 3) = F(3, 7)$ distribution is greater than 0.20 or more), then the corresponding observation is influential. For the reading versus illumination data, the influential points are (pick the two largest D_i values above)
(*choose two!*) **2** / **8** / **10**
Notice that, for observation 10 (where $D_{10} = 1.460$)

$$P(F_{3,7} < 1.460) = 0.69$$

(2nd DISTR −E99, 1.46, 3, 7) which clearly indicates that this observation (circle one) **is** / **is not** influential.

Also notice that, for observation 2 (where $D_2 = 0.474$)

$$P(F_{3,7} < 0.474) = 0.29$$

which clearly indicates that this observation (circle one) **is** / **is not** influential but not as influential as observation 10.

3. *Influence Single i Has On Each Regression Coefficient* b_k*:* $(DFBETAS)_{k(i)}$.

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

If this measure is *large* (bigger than 1 for small to medium data sets or bigger than $\sqrt{\frac{2}{n}} = \sqrt{\frac{2}{10}} \approx 0.45$ for large data sets), than the corresponding observation is influential. For the reading versus illumination data, the influential points (look at the three $(DFBETAS)_{k(i)}$ values above)
for b_0 are (circle none, one or more) **2 / 8 / 10**
for b_1 are (circle none, one or more) **2 / 8 / 10**
for b_2 are (circle none, one or more) **2 / 8 / 10**

9.5 Multicollinearity Diagnostics–Variance Inflation Factor

SAS program: att8-9-5-read-vif

The *variance inflation factors* ($(VIF)_k$) for each of the k regression coefficients is

$$(VIF)_k = \frac{1}{1 - R_k^2},$$

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. If the largest $(VIF)_k$, among the k values associated with the k regression coefficients, is *large* (greater than ten (10)) this indicates multicollinearity is a problem.

Exercise 9.5 (Multicollinearity Diagnostics–Variance Inflation Factor)

i	1	2	3	4	5	6	7	8	9	10
illumination, X_{i1}	9	7	11	16	21	19	23	29	31	33
noise, X_{i2}	15	20	17	22	24	26	28	30	29	37
ability to read, Y	70	70	75	88	91	94	100	92	90	85

1. *A First Look at Multicollinearity*

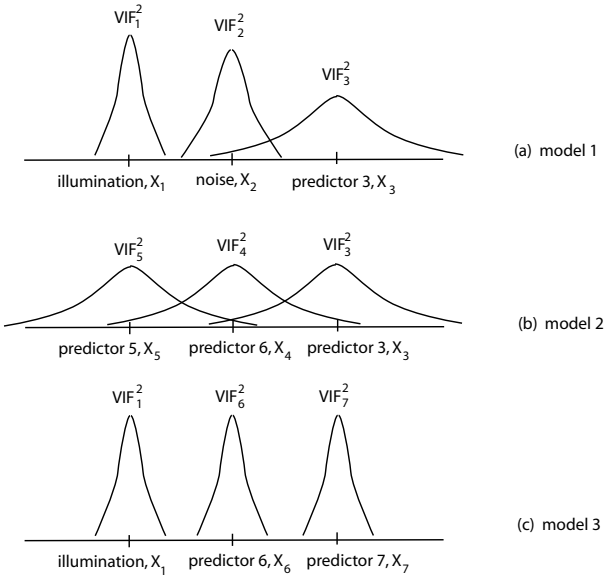


Figure 9.4 (A first look at multicollinearity)

From the figure above, the model which most clearly (unambiguously) describes the response, ability to read, Y, is (choose one)

(a) Model 1: illumination, X_1, noise, X_2 and predictor 3, X_3

(b) Model 2: predictor 5, X_5, predictor 4, X_4 and predictor 3, X_3

(c) Model 3: illumination, X_1, predictor 6, X_6 and predictor 7, X_7

Multicollinearity is a problem because, although the overall model may fit the data well, because the "variability" of several independent variables overlap one another (as measured by the $(VIF)_k$), it is difficult to decide which of the individual variables contribute significantly to the regression relationship.

2. *Calculation of Variance Inflation Factors.*
Using only the *two* predictors (illumination and noise) in the data given above, from SAS we find the variance inflation factors for the two regression coefficients in the model to both be $VIF_k = 7.25363$, $k = 1, 2$, which indicates the multicollinearity (circle one) **to be a problem / to not be a problem** since both are less than ten (10).

3. *Understanding the Variance Inflation Factor.*
   True / False
   On the one hand, the denominator of *VIF*

$$\sigma^2\{b'_k\} = \frac{(\sigma')^2}{1 - R_k^2} = (\sigma')^2(\,VIF)_k$$

   is $1 - R_k^2$, where R_k^2 is related to the correlation for the predictor variables. Since large correlation implies multicollinearity, *small* $1 - R_k^2$ also implies multicollinearity and so *large* $\frac{1}{1-R_k^2}$ implies multicollinearity.

   On the other hand, the $(\sigma')^2$ is a measure of the variance of the *standardized* regression coefficients.

   In other words, the variance inflation factor $(\,VIF)_k$ measures the factor ("inflation") amount the variance of the *non*standardized regression is larger than variance of the *standardized* regression coefficients.

4. *Informal Ways Of Investigating Multicollinearity*
   Informal ways of deciding whether multicollinearity exists in a regression model are (choose none, one or more)

   (a) large changes occur in estimated regression coefficients when predictor variable or observation is added or deleted

   (b) non–significant regression coefficients of important predictor variables

   (c) regression coefficients with the sign opposite of what it should be

   (d) large values in $\mathbf{r}_{XX}$

   (e) wide CI of regression coefficients of important predictor variables

9.6 Surgical Unit Example

An interesting example.