Chapter 10: More diagnostics

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

PRESS_p criterion

$$PRESS_p = \sum_{i=1}^{n}(Y_i - \hat{Y}_{i(i)})^2 \left(= \sum_{i=1}^{n} \left[\frac{e_i}{1 - h_{ii}} \right]^2 \right),$$

where $\hat{Y}_{i(i)}$ is the fitted value at $\mathbf{x}_i$ with the $(\mathbf{x}_i, Y_i)$ omitted.

- This is leave-one-out prediction error. The smaller, the better.
- Having $\text{PRESS}_p \approx SSE_p$ supports the *validity* of the model with p predictors (p. 374). Note that always $\text{PRESS}_p > SSE_p$, but when they're (reasonably) close, that means that there are not just a handful of points driving all inference.

9.5 Caveats for automated procedures

- `proc reg` can give you the the, say, three best subsets according to C_p containing one variable, two variables, etc. Need to define interactions & quadratic terms by hand. Cannot do it heirarchically. Best to do when number of predictors is small to moderate.
- `proc glmselect` does a great job with stepwise procedures but cannot do best subsets. Good to use when there's lots of predictors.
- There is no "best" way to search for good models.
- There may be *several* "good" models.
- If you use the same data to *estimate* the model and *choose* the model, the regression effects are *biased*!
- This leads to the idea of data splitting; one portion of the data is the *training data* and the other portion is the *validation* set (Section 9.6, p. 372). PRESS$_p$ can also be used.

Diagnostics we have already discussed

- Residuals e_i vs. each $x_1, \ldots, x_k$ and e_i vs. $\hat{Y}_i$.
- Normal probability plot of $e_1, \ldots, e_n$.
- Y_i vs. $\hat{Y}_i$. What to look for?
- VIF_j for $j = 1, \ldots, k$.
- Now we'll discuss added variable plots, leverages, dffits, and Cook's distance.

10.1 Added variable plots

- Residuals e_i versus predictors can show whether a predictor may need to be transformed or whether we should add a quadratic term.
- We can omit the predictor from the model and plot the residuals e_i versus the predictor to see if the predictor explains residual variability. Your book suggests doing this for interactions.
- An added variable plot refines this idea.
- Answers question: Does x_j explain any residual variability once the rest of the predictors are in the model?

10.1 Added variable plots

- Consider a pool of predictors $x_1, \ldots, x_k$. Let's consider predictor x_j where $j = 1, \ldots, k$.
- Regress Y_i vs. all predictors *except* x_j, call the residuals $e_i(Y|\mathbf{x}_{-j})$.
- Regress x_j vs. all predictors *except* x_j, call the residuals $e_i(x_j|\mathbf{x}_{-j})$.
- The *added variable plot* for x_j is $e_i(Y|\mathbf{x}_{-j})$ vs. $e_i(x_j|\mathbf{x}_{-j})$.
- The least squares estimate b_j obtained from fitting a line (through the origin) to the plot *is the same* as one would get from fitting the full model $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots \beta_k x_{ik} + \epsilon_i$ (Christensen, 1996).
- Gives an idea of the functional form of x_j: a transformation in x_j should mimic the pattern seen in the plot; the methods of Section 3.9 apply.

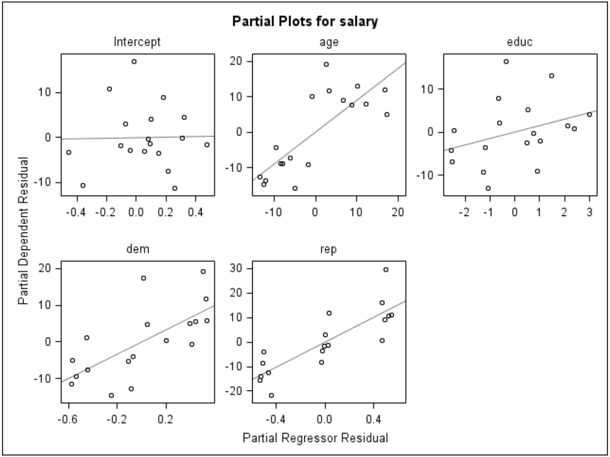
Salary data, main effects only

Partial residual plots are only in `proc reg` so need to create dummies for political affiliation.

```
data salary;
input salary age educ pol$ @@; dem=0; rep=0; if pol='D' then dem=1; if pol='R' then rep=1;
datalines;
38 25 4 D 45 27 4 R 28 26 4 O 55 39 4 D 74 42 4 R 43 41 4 O
47 25 6 D 55 26 6 R 40 29 6 O 65 40 6 D 89 41 6 R 56 42 6 O
56 32 8 D 65 33 8 R 45 35 9 O 75 39 8 D 95 65 9 R 67 69 10 O
;
ods graphics on;
proc reg;
 model salary=age educ dem rep / partial; run;
ods graphics off;
-----------------------------------------------------------------------------------------
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.49091	8.17996	0.06	0.9531
age	1	0.89835	0.19677	4.57	0.0005
educ	1	1.50395	1.18415	1.27	0.2263
dem	1	16.54042	4.88073	3.39	0.0048
rep	1	25.69912	4.75121	5.41	0.0001

Partial residual plots



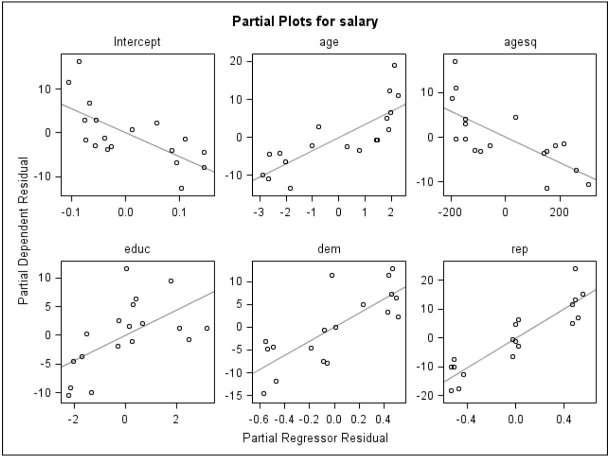
Age effect is nonlinear; let's add a quadratic term.

Salary data, quadratic effect in age

```
data salary;
input salary age educ pol$ @@; dem=0; rep=0; if pol='D' then dem=1; if pol='R' then rep=1;
 agesq=age*age;
datalines;
38 25 4 D 45 27 4 R 28 26 4 O 55 39 4 D 74 42 4 R 43 41 4 O
47 25 6 D 55 26 6 R 40 29 6 O 65 40 6 D 89 41 6 R 56 42 6 O
56 32 8 D 65 33 8 R 45 35 9 O 75 39 8 D 95 65 9 R 67 69 10 O
;
ods graphics on;
proc reg;
 model salary=age agesq educ dem rep / partial; run;
ods graphics off;
------------------------------------------------------------------------------------------
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-54.67928	16.72601	-3.27	0.0067
age	1	3.46372	0.74067	4.68	0.0005
agesq	1	-0.02883	0.00817	-3.53	0.0041
educ	1	2.16648	0.88337	2.45	0.0305
dem	1	15.45511	3.57115	4.33	0.0010
rep	1	25.57325	3.46366	7.38	<.0001

Partial residual plots w/ quadratic age



Now education is nonlinear, *but it is now significant*! The incorrect functional form for age (the effect levels off) was *masking* the importance of education.

Salary data, quadratic effect in age

```
data salary;
input salary age educ pol$ @@; dem=0; rep=0; if pol='D' then dem=1; if pol='R' then rep=1;
 agesq=age*age; educsq=educ*educ;
datalines;
38 25 4 D 45 27 4 R 28 26 4 O 55 39 4 D 74 42 4 R 43 41 4 O
47 25 6 D 55 26 6 R 40 29 6 O 65 40 6 D 89 41 6 R 56 42 6 O
56 32 8 D 65 33 8 R 45 35 9 O 75 39 8 D 95 65 9 R 67 69 10 O
;
ods graphics on;
proc reg;
 model salary=age agesq educ educsq dem rep / partial; run;
ods graphics off;
------------------------------------------------------------------------------------------
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-89.95426	17.86654	-5.03	0.0004
age	1	2.78703	0.62615	4.45	0.0010
agesq	1	-0.01868	0.00730	-2.56	0.0266
educ	1	18.75132	5.73911	3.27	0.0075
educsq	1	-1.34234	0.46111	-2.91	0.0142
dem	1	13.97691	2.84888	4.91	0.0005
rep	1	23.47204	2.81306	8.34	<.0001

Question: what amount of education is “optimal?”

Outliers

- Outliers are bizarre data points. Observations may be outlying relative only to other predictors $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})'$ or relative to *the model*, i.e. Y_i relative to $\hat{Y}_i$.
- *Studentized deleted residuals* are designed to detect outlying Y_i observations; *leverages* detect outlying $\mathbf{x}_i$ points.
- Outliers have the potential to influence the fitted regression function; they may *strengthen* inference and reduce error in predictions if the outlying points follow the modeling assumptions and are representative.
- If not, outlying values may skew inference unduly and yield models with poor predictive properties.

Outliers & influential points

- Often outliers are "flagged" and deemed suspect as mistakes or observations not gathered from the same population as the other observations.
- Sometimes outliers are of interest in their own right and may illustrate aspects of a data set that bear closer scrutiny.
- Although an observation may be flagged as an outlier, the point *may or may not* affect the fitted regression function more than other points.
- A *DFFIT* is a measure of influence that an individual point $(\mathbf{x}_i, Y_i)$ has on the regression surface at $\mathbf{x}_i$.
- *Cook's distance* is a consolidated measure of influence the point $(\mathbf{x}_i, Y_i)$ has on the regression surface at all n points $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

10.2 Studentized deleted residuals

- The *standardized residuals*

$$r_i = \frac{Y_i - \hat{Y}_i}{\sqrt{MSE(1 - h_{ii})}}$$

  have a constant variance of 1.
- Typically, $|r_i| > 2$ is considered "large." $h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ is the i^{th} leverage value.
- A refinement of the standardized residual that has a recognizable distribution is the *studentized deleted residual*

$$t_i = r_i \sqrt{\frac{MSE}{MSE_{(i)}}}$$

  where $MSE_{(i)}$ is the mean squared error calculated from a multiple regression with the same predictors but the i^{th} observation removed.
- The studentized deleted residual t_i will be larger that a regular studentized residual r_i if and only if $MSE_{(i)} < MSE$.

Studentized deleted residuals

- Recall that *MSE* is an estimated of the error variance σ^2; if including the point $(\mathbf{x}_i, Y_i)$ in the analysis increases our estimate of σ^2, then the deleted residual will be larger than the regular residual.
- Studentized deleted residuals have a computationally convenient formula (in your book) and are distributed

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

- Therefore, outlying Y-values may be flagged by using Bonferroni's adjustment and taking

$$|t_i| > t(1 - \alpha/(2n); n - p - 1)$$

  as outlying.
- Typically, in practice, one simply flags observations with $|t_i|$ larger than $t(1 - \alpha/2; n - p - 1)$ as possibly outlying in consideration with other diagnostics to be discussed shortly.

10.3 Leverage

- The leverages h_{ii} get larger the further the points $\mathbf{x}_i$ are from the mean $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$, adjusted for "how many" other predictors are in the vicinity of $\mathbf{x}_i$.
- We may use the fact that $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}\mathbf{H}$ to show $\sum_{i=1}^{n} h_{ii} = p$ and $0 \le h_{ii} \le 1$.
- A large leverage h_{ii} indicates that $\mathbf{x}_i$ is far away from the other predictors $\mathbf{x}_j$, $j \neq i$ and that $\mathbf{x}_i$ may influence the fitted value $\hat{Y}_i$ more than other $\mathbf{x}_j$'s will influence their respective fitted values. This is evident in the variance of the residual $var(Y_i - \hat{Y}_i) = \sigma^2\sqrt{(1 - h_{ii})}$. The larger h_{ii} is, the smaller $var(Y_i - \hat{Y}_i)$ will be and hence the closer $\hat{Y}_i$ will be to Y_i on average.
- The rule of thumb is that any leverage h_{ii} that is larger than twice the mean leverage p/n, i.e. $h_{ii} > 2p/n$, is flagged as having "high" leverage.

Leverage

- Note that the leverages h_{ii} depend only on the $\mathbf{x}_i$ and hence indicate which points might *potentially* be influential.
- (p. 400) When making predictions $\mathbf{x}_{n+1}$ at a point not in the data set, we consider the measure of distance of this point from the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ given by $h_{n+1} = \mathbf{x}'_{n+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{n+1}$.
- If h_{n+1} is much larger than any of the $\{h_{11}, \ldots, h_{nn}\}$ you may be extrapolating far outside the general region of your data.
- Just include an empty response (a period) in the data, but with the $\mathbf{x}_{n+1}$ information. SAS will give you h_{n+1} along with the other leverages.

10.4 DFFITS

- The i^{th} *DFFIT*, denoted $DFFIT_i$, is given by

$$DFFIT_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}},$$

  where $\hat{Y}_i$ is fitted value of regression surface (calculated using all n observations) at $\mathbf{x}_i$ and $\hat{Y}_{j(i)}$ is fitted value of regression surface *omitting the point* $(\mathbf{x}_i, Y_i)$ at the point $\mathbf{x}_j$.
- $DFFIT_i$ is standardized distance between *fitted* regression surfaces *with* and *without* the point $(\mathbf{x}_i, Y_i)$.
- Rule of thumb that $DFFIT_i$ is "large" when $|DFFIT_i| > 1$ for small to medium-sized data sets and $|DFFIT_i| > 2\sqrt{p/n}$ for large data sets. We will often just note those $DFFIT_i$'s that are considerably larger than the bulk of the $DFFIT_i$'s.

10.4 Cook's distance

- The i^{th} Cook's distance, denoted D_i, is an aggregate measure of the influence of the i^{th} observation on all n fitted values:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_i - \hat{Y}_{j(i)})^2}{p(MSE)}.$$

- This is the sum of squared distances, at each $\mathbf{x}_j$, between fitted regression surface calculated with all n points and fitted regression surface calculated with the i^{th} case removed, standardized by $p(MSE)$.
- Look for values of Cook's distance significantly larger than other values; these are cases that exert disproportionate influence on the fitted regression surface as a whole.

Review of diagnostics

- **Variance inflation factors** VIF_j tell you which predictors are highly correlated with other predictors. If you have one or more $\mathrm{VIF}_j > 10$, you may want to eliminate some of the predictors.

  Multicollinearity affects the interpretation of the model, but does not indicate the model is "bad" in any way.

  An alternative approach that allows keeping correlated predictors is ridge regression (Chapter 11).
- **Deleted residuals** $t_i \sim t_{n-p-1}$, so you can formally define an outlier as being larger than $t_{n-p-1}(1 - \alpha/(2n))$.

Review of diagnostics

- **Residual plots**. Plots of e_i or t_i vs. $\hat{Y}_i$ and versus each $x_1, \ldots, x_k$ help assess (a) correct functional form, (b) constant variance, and (c) outlying observations. If an anomaly is apparent in any of these plots I may look at an added variable plot. If the number of predictors is small I may look at every added variable plot. These plots indicate problems such as non-constant variance and the appropriateness of a plane as a regression surface. They may also suggest a transformation for a predictor or two.
  - Heteroscedasticy can be corrected by transforming Y, or else modeling the variance directly (Chapter 11).
  - Constant variance but nonlinear patterns can be accommodated by introducing quadratic terms.
- **Added variable plots** help figure out functional form of predictors, and whether significance is being driven by one or two points only.
- `proc transreg` and `proc gam` fit models where every predictor can be transformed simultaneously.

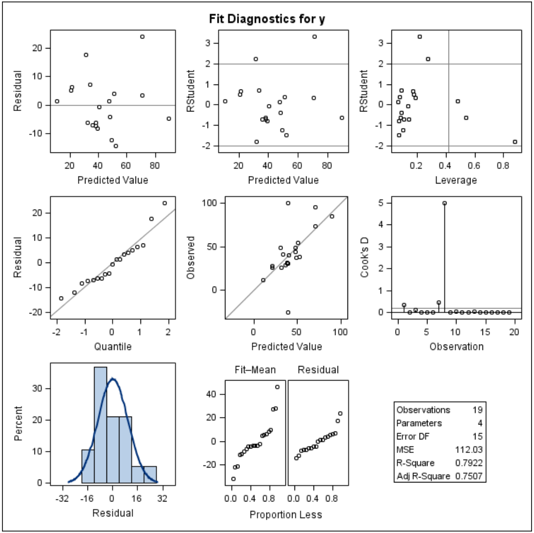
Review of diagnostics

- **DFFIT**$_i$ and **Cook's distance** D_i tell you which observations influence the fitted model the most. Sometimes one or two points can drive the significance of an effect.
- **Leverages** tell you which points *can potentially* influence the fitted model. Useful for finding "hidden extrapolations" via h_{n+1}.
- (pp. 404–405) DFBETA$_{ij}$ tells you how much observation i affects regression coefficient j. Useful to "zoom in" on where influential points are affecting the model.
- A **normal probability plot** of the residuals will indicate gross departures from normality.
- A list of the studentized deleted residuals, leverages, and Cook's distances helps to determine outlying values that may be transcription errors or data anomalies and also indicates those observations that affect the fitted regression surface as a whole.

Standard SAS diagnostic plots

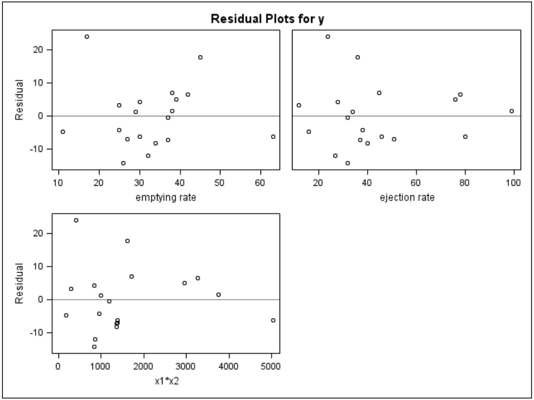
- t_i vs. h_i. Which observations are outlying in $\mathbf{x}$-direction, outlying in Y-direction, or both?
- D_i vs. i. Which observations grossly affect fit of regression surface?
- e_i vs. $\hat{Y}_i$ and t_i vs. $\hat{Y}_i$. Constant variance & linearity.
- Y_i vs. $\hat{Y}_i$; how well model predicts its own data. Better models have points close to line $y = x$.
- Normal probability plot of the $e_1, \ldots, e_n$.
- Histogram of $e_1, \ldots, e_n$.
- Plots of e_i vs. each predictor $x_1, \ldots, x_k$.
- One more plot that I never look at.

Arterial pressure data in `proc glm`



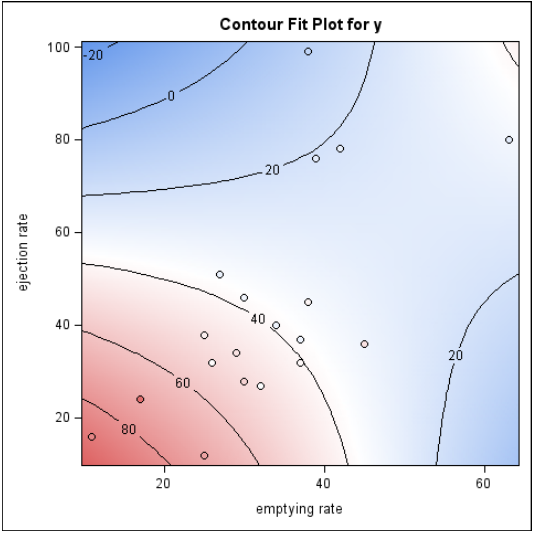
Model is $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \epsilon_i$. One highly influential point & one poorly fit.

Arterial pressure data in proc glm



These look pretty good, aside from the one large residual.

Arterial pressure data in `proc glm`



`proc glm` recognizes that there are only two variables and plots a response surface automatically.

Arterial pressure data in `proc glm`

```
proc glm;
 model y=x1 x2 x1*x2;
 output out=out cookd=c rstudent=r; run;
proc print; var x1 x2 y c r; run;
--------------------------------------------------------------------
Obs     x1     x2     y        c           r
  1     45     36     49     0.36904     2.20950
  2     30     28     55     0.00383     0.39889
  3     11     16     85     0.12052    -0.62921
  4     30     46     32     0.00885    -0.60493
  5     39     76     26     0.01498     0.51721
  6     42     78     28     0.02392     0.66178
  7     17     24     95     0.45892     3.31414
  8     63     80     26     4.99081    -1.77941
  9     25     12     74     0.00724     0.33794
 10     32     27     37     0.04100    -1.22324
 11     37     37     31     0.01660    -0.71526
 12     29     34     49     0.00032     0.12816
 13     26     32     38     0.04023    -1.45743
 14     38     45     41     0.01271     0.69211
 15     38     99     12     0.00817     0.18206
 16     25     38     44     0.00422    -0.40213
 17     27     51     29     0.02196    -0.70921
 18     37     32     40     0.00014    -0.05730
 19     34     40     31     0.01371    -0.80210
```

Obs. 7 has largest arterial pressure. Obs. 8 has relatively small arterial pressure.

Two subsets

```
proc glm data=out; model y=x1 x2 x1*x2; run;
-----------------------------------------------------------------
                                  Standard
Parameter         Estimate           Error    t Value    Pr > |t|
Intercept      134.3998664     15.98159869       8.41      <.0001
x1              -2.1330220      0.52215739      -4.09      0.0010
x2              -1.6993299      0.36366865      -4.67      0.0003
x1*x2            0.0333471      0.00928281       3.59      0.0027
-----------------------------------------------------------------
proc glm data=out(where=(c<4)); model y=x1 x2 x1*x2; run;
-----------------------------------------------------------------
                                  Standard
Parameter         Estimate           Error    t Value    Pr > |t|
Intercept      157.5094488     19.79515582       7.96      <.0001
x1              -2.7122125      0.58667658      -4.62      0.0004
x2              -2.7743376      0.69321545      -4.00      0.0013
x1*x2            0.0618590      0.01822201       3.39      0.0044
-----------------------------------------------------------------
proc glm data=out(where=(abs(r)<3)); model y=x1 x2 x1*x2; run;
-----------------------------------------------------------------
                                  Standard
Parameter         Estimate           Error    t Value    Pr > |t|
Intercept      116.3928224     13.52293668       8.61      <.0001
x1              -1.6161083      0.43361763      -3.73      0.0023
x2              -1.4903775      0.28875668      -5.16      0.0001
x1*x2            0.0272510      0.00742428       3.67      0.0025
```

How do 7 and 8 affect the significance and/or magnitude of the effects?

More examples...

- Surgical unit data.
- Salary data.
- Body fat data.