References

- G&S Chapter 4: Do the data fit the assumptions?
- KKMN Chapter 12: Regression diagnostics
- NKNW Chapter 9

3 Sets of issues...

1 At the level of the individual observation...

- outliers in X ---> "influence"
- outliers in Y ---> residuals, rough & refined
- overall effect ---> Cooks' "distance"
  ---> Δ Fit
  ---> Δ β's

2 At the level of the model

- model mis-specified
  - variables omitted
  - have important variables, but mis-specify form of equation

Analysis of Residuals key in -1- and -2-

-3- At the level of the variables

- joint distribution of variables less than optimal
  (multi)collinearity (Chapter 5)

1. AT LEVEL OF THE INDIVIDUAL OBSERVATION...

Potential to Influence fitted model

(see G&S Chapter 4 pages 130-134)

- Leverage (h)
  - function of x's (cf. [$X - \bar{x}$] term in SE for projection
  - $0 < h \leq 1$;

    average value: $\frac{\#\ \textit{terms} \text{ in equation}}{n}$ [# includes β_0]

    "watch out for" value: > 2 × **average** value

Residuals

(see G&S Chapter 4 pages 119-130 and 134-136)

- *unscaled* ("raw"): $e = y - \hat{y}$

- *standardized:* e / RMSE

- *refined*: e's at "X-edges" are less variable: $\text{var}(e) = \sigma^2(1 - h)$

  studentized:

  $e / \{ \text{RMSE} \sqrt{1 - h} \}$

- *refined further* to make outliers "stand out",
  re-estimate RMSE by deleting observation

  studentized deleted:

  $e / \{ \text{RMSE}_{(-i)} \sqrt{1 - h} \}$

1. AT LEVEL OF THE INDIVIDUAL OBSERVATION...

Residuals (see comment re studentized residuals at top of p 135 of G&S)

G&S text name	G&S text symbol	definition	SAS INSIGHT name	SAS INSIGHT prefix	SAS Proc REG Keywords in Output and Plot subcommands	NKNW text
unscaled / "raw"	e	$y - \hat{y}$	residual	**R_Y**	residual	e
standardized	e_s	$\frac{y - \hat{y}}{RMSE}$				e*
studentized/ internally standardized	r	$\frac{y - \hat{y}}{RMSE\sqrt{1 - h}}$	standardized	**RS_Y**	student	r
studentized deleted / jackknife/ externally standardized	$r_{(-i)}$	$\frac{y - \hat{y}}{RMSE_{(-i)}\sqrt{1 - h}}$	studentized	**RT_Y**	rstudent	t

1. AT LEVEL OF THE INDIVIDUAL OBSERVATION...

Actual Influence on fitted model

Cook's Distance (D)

- (Scaled) Distance of regression coefficients [b's] obtained without the observation from those obtained with all n observations.

- Also expressible as..
(Scaled) Distance of fitted y's obtained without the observation from those obtained with all n observations.

- D's "tend to follow" $F_{(\#\ terms, n\ -\ \#\ terms)}$ distribution, so

D > 1 **D > 4**

see also Ch 232 of KKMN

- D is a function of studentized residual (r) and leverage (h)

$$D = \frac{r^2}{\#\ \text{of terms}} \times \frac{h}{1 - h}$$

so larger if larger r and larger h

Change in Fitted (predicted) Values (DFFitsS)

- the amount by which the predicted Y value for the observation in question changes when the observation itself is excluded from the analysis.

- a standardized measure

$$\textbf{DFFitsS} > 2\sqrt{\frac{\#\ \text{terms}}{n}}$$

Dfbeta's

- (Standardized) measure of amount by which each term in the model changes when the observation itself is excluded from the analysis.

DFbeta > 2

2. AT THE LEVEL OF THE MODEL

Model evaluation...

(see G&S Chapter 4 pages 145-151 and 170)

The right variables, in the appropriate form?

- Residual Plots

  Worry more about bimodal distributions of residuals than lack of Gaussian-ness (e.g long tails) per se. A bimodal distribution might hint at an omitted binary covariate

In <u>Simple</u> Linear Regression

- can directly identify need for quadratic term in X, other transformation etc.

In <u>Multiple</u> Linear regression...

- plot residuals vs predicted and against each X
- plot Partial Y residuals vs. partial X residuals Called **Partial Leverage** Plots in INSIGHT (cf Annotated guide**)**

3. AT THE LEVEL OF THE VARIABLES...

Multi-collinearity *cf Chapter 5 and annotated guide to Output...*

how to diagnose it

Pairwise correlations of X's

Variance Inflation Factors (better),together with collinearity diagnostics

If estimates of b's "flip" (change sign)
(remember the hammock or trampoline!!)

or have very large Standard Errors

when does Multicollinearity matter?

if seek reasonably uncorrelated estimates of 2 or more β's but unfavourable distribution of corresponding X's

e.g. if in a study of hearing (lung function) loss, wish to "separate" the β for age from the β for years worked in noisy (dusty) jobs

what to do about it

- drop one X (or drop the project!)
- get outside estimates for some of the β's
- increase sample size, and study more favourable X's

- if adding powers or products of other variables, ***center*** X variables first!

when does Multicollinearity matter less?

if merely interested in prediction even then, may want to reduce the "dimensionality" of the X's using a technique such as principal components