

Applied Regression Modeling: A Business Approach

Chapter 5: Regression Model Building II

Sections 5.1–5.2

by Iain Pardoe

5.1 Influential points	2
Influential points	2
Outliers	3
Dealing with outliers	4
CARS5 data	5
Model 1 studentized residuals	6
Remove outlier	7
Model 2 studentized residuals	8
Leverage	9
Results with outlier removed	10
Model 2 leverages	11
Results with high leverage point removed	12
Cook's distance	13
Results for all observations	14
Model 1 Cook's distances	15
Results with outlier removed	16
Model 2 Cook's distances	17
5.2 Regression pitfalls	18
Regression pitfalls	18
Autocorrelation	19
Model 1 residuals	20
Model 2 residuals	21
Multicollinearity	22
Model 1 results	23
X_1 and X_2 highly correlated	24
Model 2 results	25
Excluding important predictor variables	26

Model 1 results	27
Model 2 results	28
Paradox explained	29
Overfitting	30
Extrapolation	31
Extrapolation example	32
Missing data	33
Model results	34

Influential points

- Beware model conclusions that are overly influenced by a small handful of data points, e.g.:
 - overall results can be biased if a few unusual points differ dramatically from general patterns in the majority of the data values;
 - misleading to conclude evidence of a strong association between variables if evidence based mainly on a few dominant points.
- Focus on two measures of individual data point influence:
 - *outliers* have unusual Y -values relative to their predicted \hat{Y} -values from a model;
 - high *leverage* points have unusual combinations of X -values relative to general dataset patterns.
- *Cook's distance* is a composite measure of outlyingness and leverage.

© Iain Pardoe, 2006

2 / 34

Outliers

- Outliers have unusual Y -values relative to their predicted \hat{Y} -values from a model.
- In other words, observations with a large magnitude residual, $\hat{e}_i = Y_i - \hat{Y}_i$.
- Computer can calculate *studentized* residuals to put them on a common scale.
- When four regression assumptions (zero mean, constant variance, normality, and independence) are satisfied, studentized residuals $\approx N(0, 1^2)$.
- If we identify an observation with a studentized residual outside $(-3, 3)$, we've either witnessed a very unusual event (one with prob. less than 0.002) or we've found an observation with a Y -value that doesn't fit the pattern in the rest of the dataset.
- Formally define a potential outlier as an observation with studentized residual < -3 or > 3 .

© Iain Pardoe, 2006

3 / 34

Dealing with outliers

- If we find one or more outliers, investigate why:
 - data input mistake (remedy: identify and correct mistake(s) and reanalyze data);
 - important predictor omitted from model (remedy: identify potentially useful predictors not included in the model and reanalyze data);
 - regression assumptions violated (remedy: reformulate model using transformations or interactions, say, to correct problem);
 - potential outliers differ substantively from other sample observations (remedy: remove outliers and reanalyze remainder of dataset separately).
- To gauge outlier influence exclude largest magnitude studentized residual, refit model to remaining observations, and see if regression parameter estimates change substantially.

© Iain Pardoe, 2006

4 / 34

CARS5 data

- Y = city miles per gallon (MPG) for 50 new U.S. passenger cars in 2004.
- X_1 = weight (thousands of pounds).
- X_3 = engine size (liters).
- X_5 = wheelbase (hundreds of inches).
- Model: $E(Y) = b_0 + b_1(1/X_1) + b_2(1/X_3) + b_3(1/X_5)$.

Parameters ^a				
Model	Estimate	Std. Error	t-stat	Pr(> t)
1 (Intercept)	9.397	13.184	0.713	0.480
recipX1	44.296	13.173	3.363	0.002
recipX3	19.404	6.706	2.894	0.006
recipX5	-9.303	17.087	-0.544	0.589

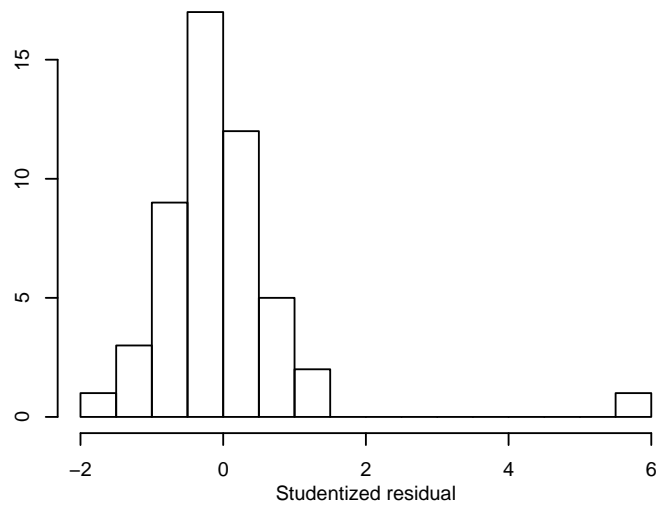
^a Response variable: Y .

© Iain Pardoe, 2006

5 / 34

Model 1 studentized residuals

There is one outlier with studentized residual ≈ 6 .



© Iain Pardoe, 2006

6 / 34

Remove outlier

Parameters ^a

Model	Estimate	Std. Error	t-stat	Pr(> t)
2 (Intercept)	25.946	7.612	3.409	0.001
recipX1	64.071	7.682	8.340	0.000
recipX3	17.825	3.782	4.713	0.000
recipX5	-33.106	9.919	-3.338	0.002

^a Response variable: Y.

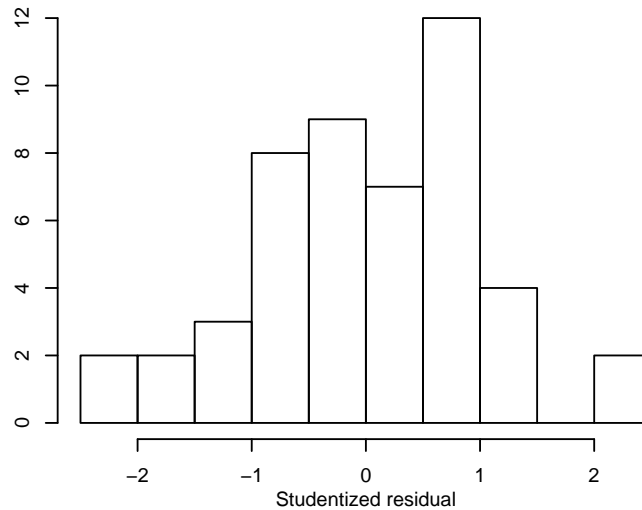
- Regression parameter estimates and p-values change dramatically.
- The outlier was diesel-powered and did not fit the pattern of the gasoline-powered cars.

© Iain Pardoe, 2006

7 / 34

Model 2 studentized residuals

No further outliers.



© Iain Pardoe, 2006

8 / 34

Leverage

- High *leverage* points have unusual combinations of X -values relative to general dataset patterns.
- If a point is far from the majority of the sample, it can pull the fitted model close toward its Y -value, potentially biasing the results.
- Leverage measures potential for an observation to have undue influence on a model (0–1: low–high).
- Rule of thumb:
 - if leverage $> 3(k + 1)/n$ investigate further;
 - if leverage $> 2(k + 1)/n$ *and isolated* investigate further;
 - otherwise, no evidence of undue influence.
- To gauge influence exclude largest leverage point, refit model to remaining observations, and see if reg. parameter estimates change substantially.

© Iain Pardoe, 2006

9 / 34

Results with outlier removed

Parameters^a

Model	Estimate	Std. Error	t-stat	Pr(> t)
2 (Intercept)	25.946	7.612	3.409	0.001
recipX1	64.071	7.682	8.340	0.000
recipX3	17.825	3.782	4.713	0.000
recipX5	-33.106	9.919	-3.338	0.002

^a Response variable: Y.

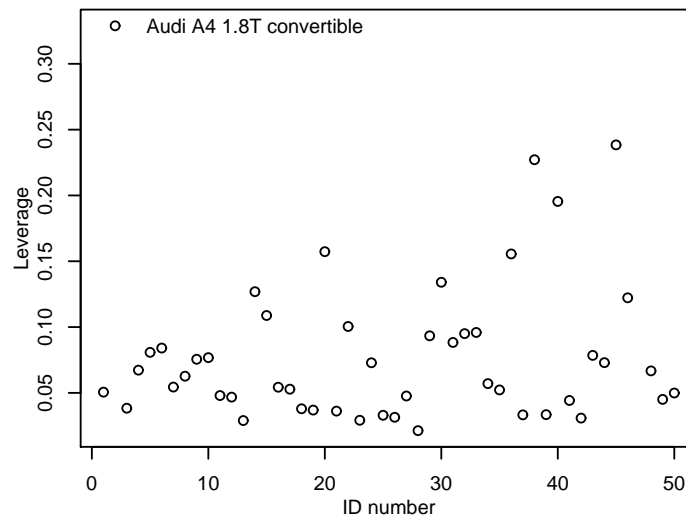
- Threshold: $3(k + 1)/n = 3(3 + 1)/49 = 0.24$.
- Threshold: $2(k + 1)/n = 2(3 + 1)/49 = 0.16$.

© Iain Pardoe, 2006

10 / 34

Model 2 leverages

Highest leverage point exceeds $3(k + 1)/n$ threshold.



© Iain Pardoe, 2006

11 / 34

Results with high leverage point removed

Parameters ^a					
Model		Estimate	Std. Error	t-stat	Pr(> t)
3	(Intercept)	24.811	7.717	3.215	0.002
	recipX1	68.054	8.785	7.747	0.000
	recipX3	15.743	4.389	3.587	0.001
	recipX5	-32.400	9.960	-3.253	0.002

^a Response variable: Y.

- Regression parameter estimates and p-values don't change dramatically.
- The high leverage point had the potential to strongly influence results, but in this case did not do so.

© Iain Pardoe, 2006

12 / 34

Cook's distance

- *Cook's distance* is a composite measure of outlyingness and leverage.
- Rule of thumb:
 - observations with a Cook's distance > 1 are *often* sufficiently influential that they should be removed from the main analysis—investigate further;
 - observations with a Cook's distance > 0.5 are *sometimes* sufficiently influential that they should be removed from the main analysis—investigate further;
 - otherwise, no evidence of undue influence.
- To gauge influence exclude largest Cook's distance, refit model to remaining observations, and see if reg. parameter estimates change substantially.

© Iain Pardoe, 2006

13 / 34

Results for all observations

Parameters ^a					
Model		Estimate	Std. Error	t-stat	Pr(> t)
1	(Intercept)	9.397	13.184	0.713	0.480
	recipX1	44.296	13.173	3.363	0.002
	recipX3	19.404	6.706	2.894	0.006
	recipX5	-9.303	17.087	-0.544	0.589

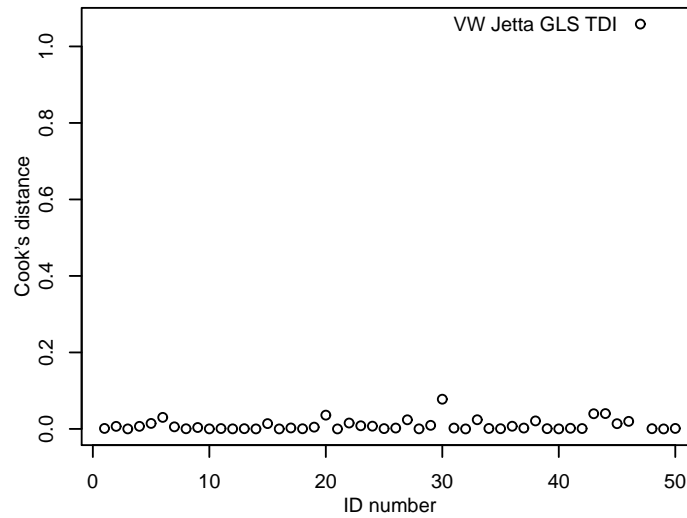
^a Response variable: Y.

© Iain Pardoe, 2006

14 / 34

Model 1 Cook's distances

Highest Cook's distance exceeds 1 threshold.



© Iain Pardoe, 2006

15 / 34

Results with outlier removed

Parameters ^a

Model	Estimate	Std. Error	t-stat	Pr(> t)
2 (Intercept)	25.946	7.612	3.409	0.001
recipX1	64.071	7.682	8.340	0.000
recipX3	17.825	3.782	4.713	0.000
recipX5	-33.106	9.919	-3.338	0.002

^a Response variable: Y.

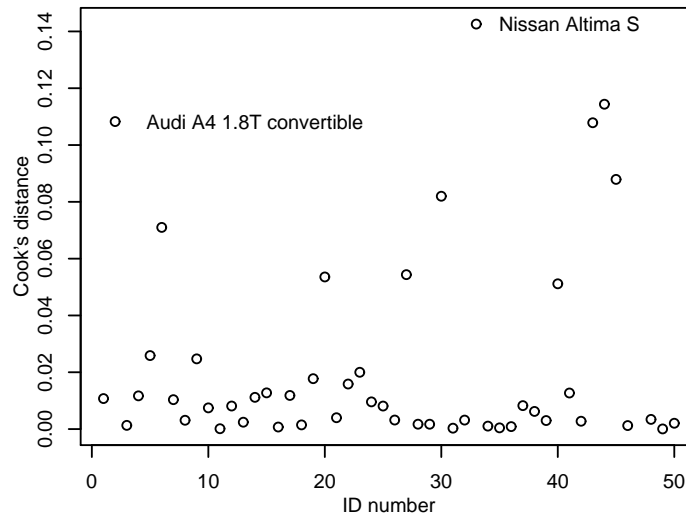
- The car with the highest Cook's distance was the outlier we found before.

© Iain Pardoe, 2006

16 / 34

Model 2 Cook's distances

Highest Cook's distance less than 0.5 threshold.



© Iain Pardoe, 2006

17 / 34

5.2 Regression pitfalls

18 / 34

Regression pitfalls

- Some of the pitfalls that can cause problems with a regression analysis:
 - autocorrelation (serial correlation)—failing to account for time trends in the model;
 - multicollinearity—highly correlated predictors causing unstable model results;
 - excluding important predictor variables—leading to possibly incorrect conclusions;
 - overfitting (the sample data)—leading to poor generalizability to the population;
 - extrapolation—using model results for predictor values very different to those in the sample;
 - missing data—leading to reduced sample sizes at best, misleading results at worst.

© Iain Pardoe, 2006

18 / 34

Autocorrelation

- Autocorrelation occurs when regression model residuals violate the independence assumption because they are highly dependent across time.
- Can occur when regression data have been collected over time and model fails to account for any strong time trends.
- Dealing with this issue rigorously can require specialized time series and forecasting methods.
- Sometimes, however, simple ideas can mitigate autocorrelation problems.
- Example: **OIL** data file contains annual world crude oil production in millions of barrels (Y) from 1880 to 1972 (X).
- Model 1: $E(\log_e(Y)) = b_0 + b_1X$.

© Iain Pardoe, 2006

19 / 34

Model 1 residuals

Plot shows clear evidence of autocorrelation.



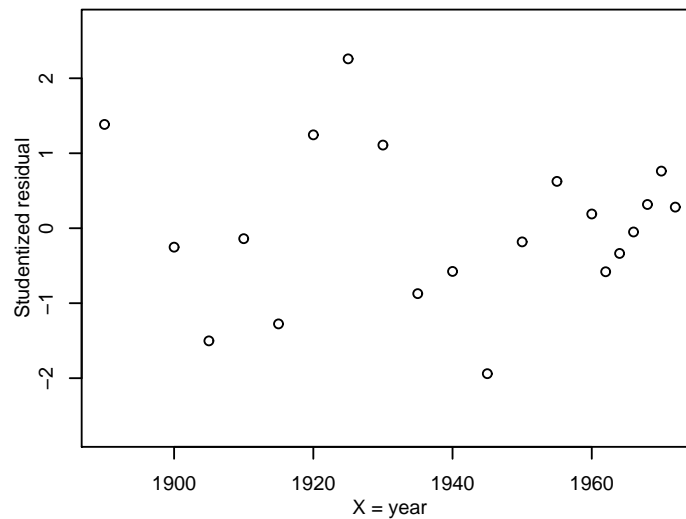
© Iain Pardoe, 2006

20 / 34

Model 2 residuals

Model 2: $E(\log_e(Y_t)) = b_0 + b_1 X_t + b_2 \log_e(Y_{t-1})$.

Independent errors assumption more reasonable now.



© Iain Pardoe, 2006

21 / 34

Multicollinearity

- Multicollinearity occurs when excessive correlation between quantitative predictors leads to unstable models and inflated standard errors.
- Identify by looking at a scatterplot matrix, calculating bivariate correlations, and calculating variance inflation factors (problem if > 10).
- Potential remedies include:
 - collect more uncorrelated data (if possible);
 - create new combined predictor variables from the highly correlated predictors (if possible);
 - remove one of the highly correlated predictors from the model.
- Example: **SALES3** data file with sales (Y), TV/newspaper advertising (X_1), and internet advertising (X_2).

© Iain Pardoe, 2006

22 / 34

Model 1 results

- Model 1: $E(Y) = b_0 + b_1X_1 + b_2X_2$.

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.987 ^a	0.974	0.968	0.8916

^a Predictors: (Intercept), X1, X2.

Parameters^a

Model		Estimate	Std. Error	t-stat	Pr(> t)	VIF
1	(Intercept)	1.992	0.902	2.210	0.054	
	X1	0.767	0.868	0.884	0.400	49.541
	X2	1.275	0.737	1.730	0.118	49.541

^a Response variable: Y.

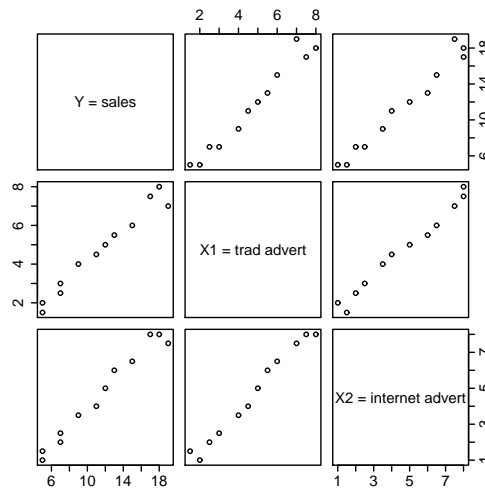
- R^2 is 0.974, but neither X_1 nor X_2 are significant (given the presence of the other)!
- VIF > 10 suggests there is a multicollinearity problem.

© Iain Pardoe, 2006

23 / 34

X_1 and X_2 highly correlated

Unstable estimates when both X_1 and X_2 in model.



© Iain Pardoe, 2006

24 / 34

Model 2 results

- Model 2: $E(Y) = b_0 + b_1(X_1 + X_2)$.

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
2	0.987 ^a	0.974	0.971	0.8505

^a Predictors: (Intercept), X1plusX2.

Parameters^a

Model	Estimate	Std. Error	t-stat	Pr(> t)
2 (Intercept)	1.776	0.562	3.160	0.010
X1plusX2	1.042	0.054	19.240	0.000

^a Response variable: Y.

- R^2 unchanged, and the combined predictor variable, $X_1 + X_2$, is significant.
- Note this approach is only possible if it makes sense to create a combined predictor variable.
- More common to drop one of the correlated predictors from model.

© Iain Pardoe, 2006

25 / 34

Excluding important predictor variables

- Excluding important predictors sometimes results in models that provide incorrect, biased conclusions about included predictors.
- Strive to include all potentially important predictors, and remove a predictor only if there are compelling reasons to do so (e.g., if causing multicollinearity problems and has high individual p-value).
- Example: **PARADOX** data file with $n=27$ high-precision computer components with component quality (Y) potentially depending on two controllable machine factors, speed (X_1) and angle (X_2).

© Iain Pardoe, 2006

26 / 34

Model 1 results

- Model 1: $E(Y) = b_0 + b_1X_1$.

Parameters ^a					
Model		Estimate	Std. Error	t-stat	Pr(> t)
1	(Intercept)	2.847	1.011	2.817	0.009
	X1	0.430	0.188	2.288	0.031

^a Response variable: Y.

- Results suggest a positive association between quality and speed.
- In other words, increase the speed of the machine to improve quality.
- However, this ignores process information relating to angle.

© Iain Pardoe, 2006

27 / 34

Model 2 results

- Model 2: $E(Y) = b_0 + b_1X_1 + b_2X_2$.

Parameters ^a					
Model		Estimate	Std. Error	t-stat	Pr(> t)
2	(Intercept)	1.638	0.217	7.551	0.000
	X1	-0.962	0.071	-13.539	0.000
	X2	2.014	0.086	23.473	0.000

^a Response variable: Y.

- Results suggest a *negative* association between quality and speed (for a fixed angle),
- and a positive association between quality and angle (for a fixed speed).
- In other words, increase the angle but decrease the speed of the machine to improve quality.

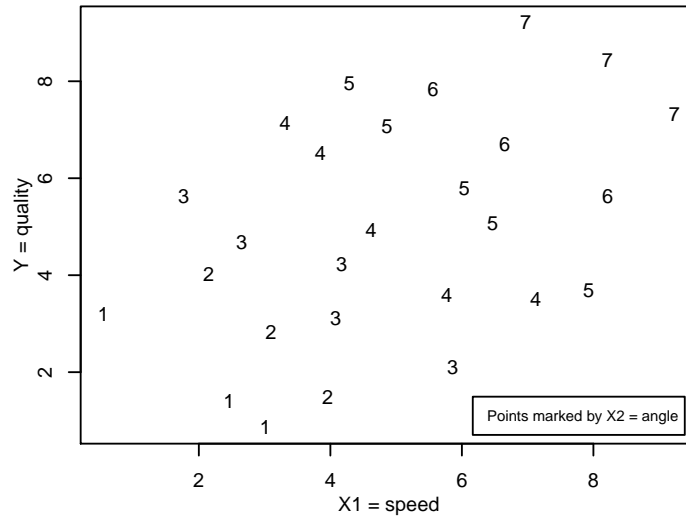
© Iain Pardoe, 2006

28 / 34

Paradox explained

Positive association between Y and X_1 ignoring X_2 .

Negative association between Y and X_1 accounting for X_2 .

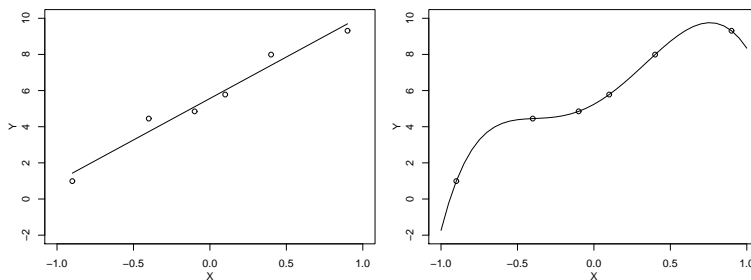


© Iain Pardoe, 2006

29 / 34

Overfitting

- Overfitting can occur if overly complicated model tries to account for every possible pattern in sample data, but generalizes poorly to underlying population.
- Should always apply a “sanity check” to make sure model makes sense from subject-matter perspective and conclusions are supported by data.



- Which model seems more reasonable?

© Iain Pardoe, 2006

30 / 34

Extrapolation

- Extrapolation occurs when regression model results are used to estimate or predict a response value for an observation with predictor values that are very different from those in the sample.
- This can be dangerous because it means making a decision about a situation where there are no data values to support our conclusions.
- Example: if we observe an upward trend between two variables, should we assume the trend continues indefinitely at higher values?

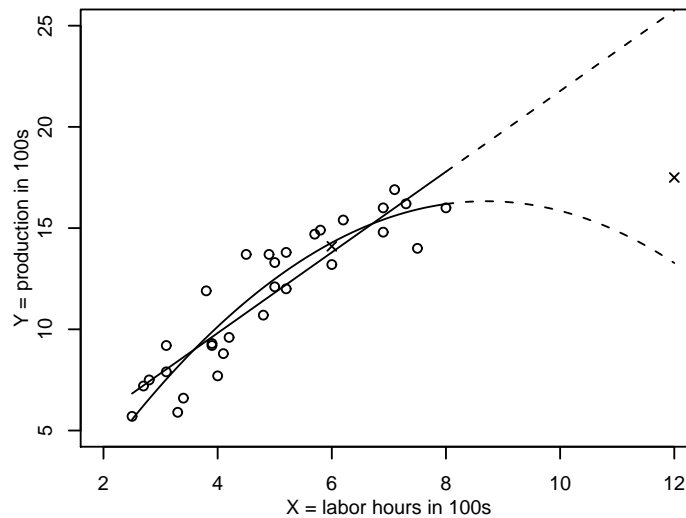
© Iain Pardoe, 2006

31 / 34

Extrapolation example

Straight-line model overshoots actual Y at far right.

Quadratic model undershoots (i.e., neither model enables accurate prediction far from sample data).



© Iain Pardoe, 2006

32 / 34

Missing data

- Missing data occurs when particular values in the dataset have not been recorded for particular variables and observations.
- Dealing with issue rigorously is beyond scope of book, but there are some simple ideas that can mitigate some of the major problems.
- Example: **MISSING** data file with $n=30$, Y , and X_1-X_4 .
- No missing values for Y , X_1 , or X_4 , but five missing values for X_2 , one of which is also missing X_3 :
 - any model including X_2 will exclude five observations;
 - including X_3 (but excluding X_2) will exclude one observation;
 - excluding X_2 and X_3 will exclude no observations.

© Iain Pardoe, 2006

33 / 34

Model results

Predictors	Sample size	R^2	s
X_1, X_2, X_3, X_4	25	0.959	0.865
X_2, X_3, X_4	25	0.958	0.849
X_1, X_3, X_4	29	0.953	0.852
X_1, X_4	30	0.640	2.300

- Ordinarily, we would probably favor the (X_2, X_3, X_4) model.
- However, the (X_1, X_3, X_4) model applies to much more of the sample.
- Thus, in this case, we would probably favor the (X_1, X_3, X_4) model, since R^2 and s are roughly equivalent, but the usable sample size is much larger.

© Iain Pardoe, 2006

34 / 34