Regression diagnostics

Biometry 755

Spring 2009

Introduction

Every statistical method is developed based on assumptions. The validity of results derived from a given method depends on how well the model assumptions are met. Many statistical procedures are "robust", which means that only extreme violations from the assumptions impair the ability to draw valid conclusions. Linear regression falls in the category of robust statistical methods. However, this does not relieve the investigator from the burden of verifying that the model assumptions are met, or at least, not grossly violated. In addition, it is always important to demonstrate how well the model fits the observed data, and this is assessed in part based on the techniques we'll learn in this lecture.

Different types of residuals

Recall that the residuals in regression are defined as $y_i - \hat{y}_i$, where y_i is the observed response for the ith observation, and $\hat{y}_i$ is the fitted response at x_i.

There are other types of residuals that will be useful in our discussion of regression diagnostics. We define them on the following slide.

Different types of residuals (cont.)

Raw residuals: $r_i = y_i - \hat{y}_i$

Standardized residuals: $z_i = \frac{r_i}{s}$ where s is the estimated error standard deviation (i.e. $s = \hat{\sigma} = \sqrt{\text{MSE}}$).

Studentized residuals: $r_i^* = \frac{z_i}{\sqrt{1-h_i}}$ where h_i is called the *leverage*. (More later about the interpretation of h_i.)

Jackknife residuals: $r_{(-i)} = r_i^* \frac{s}{s_{(-i)}}$ where $s_{(-i)}$ is the estimated error standard deviation computed with the ith observation deleted.

Which residual to use?

The standardized, studentized and jackknife residuals are all scale independent and are therefore preferred to raw residuals. Of these, jackknife residuals are most sensitive to outlier detection and are superior in terms of revealing other problems with the data. For that reason, most diagnostics rely upon the use of jackknife residuals. Whenever we have a choice in the residual analysis, we will select jackknife residuals.

Analysis of residuals - Normality

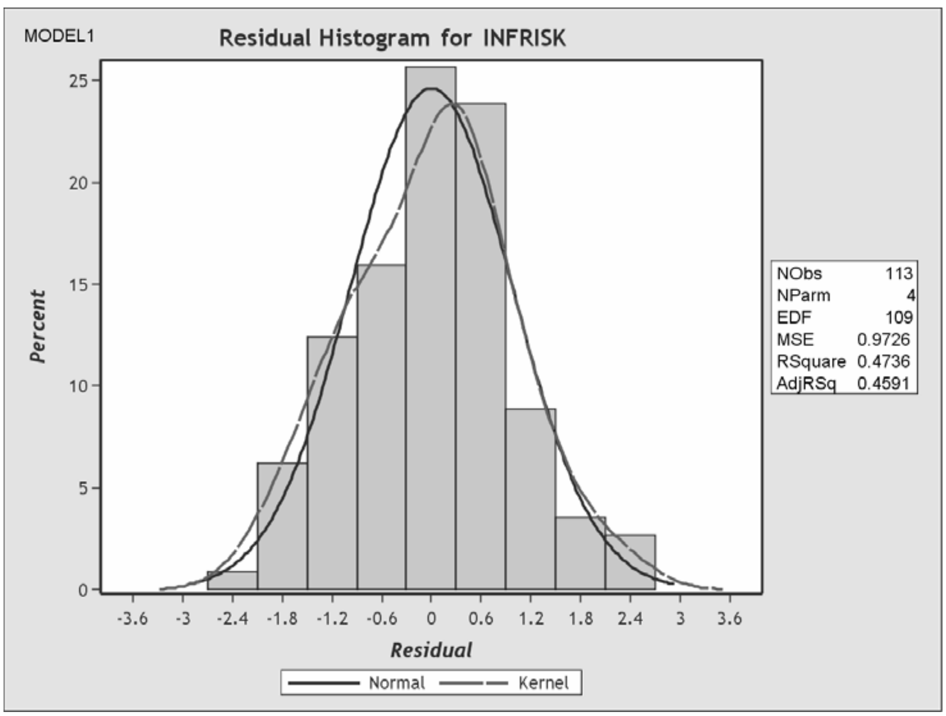
Recall that an assumption of linear regression is that the error terms are normally distributed. That is $\varepsilon \sim$ Normal$(0, \sigma^2)$. To assess this assumption, we will use the residuals to look at:

- histograms
- normal quantile-quantile (qq) plots
- Wilk-Shapiro test

Histogram and QQ plot of residuals in SAS

```
ods rtf style = Analysis;
ods graphics on;

ods select ResidualHistogram;
ods select QQPlot;

proc reg data = one plots(unpack);
    model infrisk = los cult beds;
run;
quit;

ods graphics off;
ods rtf close;
```


Histogram of residuals in SAS



What is a normal QQ plot?

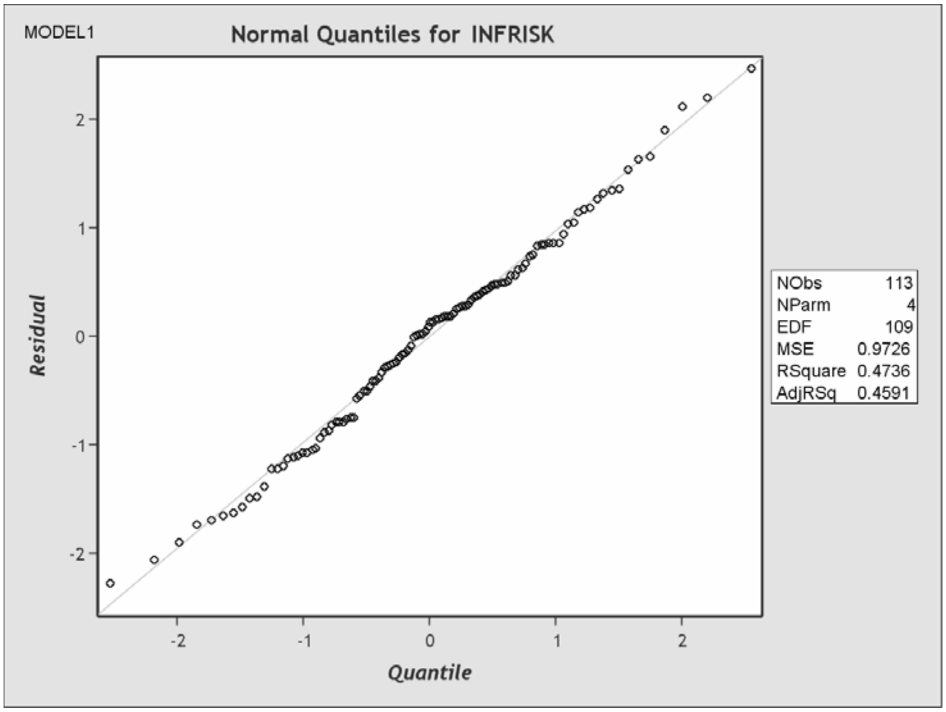
- Let q be a number between 0 and 1. The qth quantile of a distribution is that point, x, at which $q \times 100$ percent of the data lie below x and $(1-q) \times 100$ percent of the data lie above x. Specially named quantiles include quartiles, deciles, etc.
- The quantiles of the standard normal distribution are well known. Here are a few with which you should be familiar.

q	Quantile
0.025	-1.96
0.05	-1.645
0.5	0
0.95	1.645
0.975	1.96

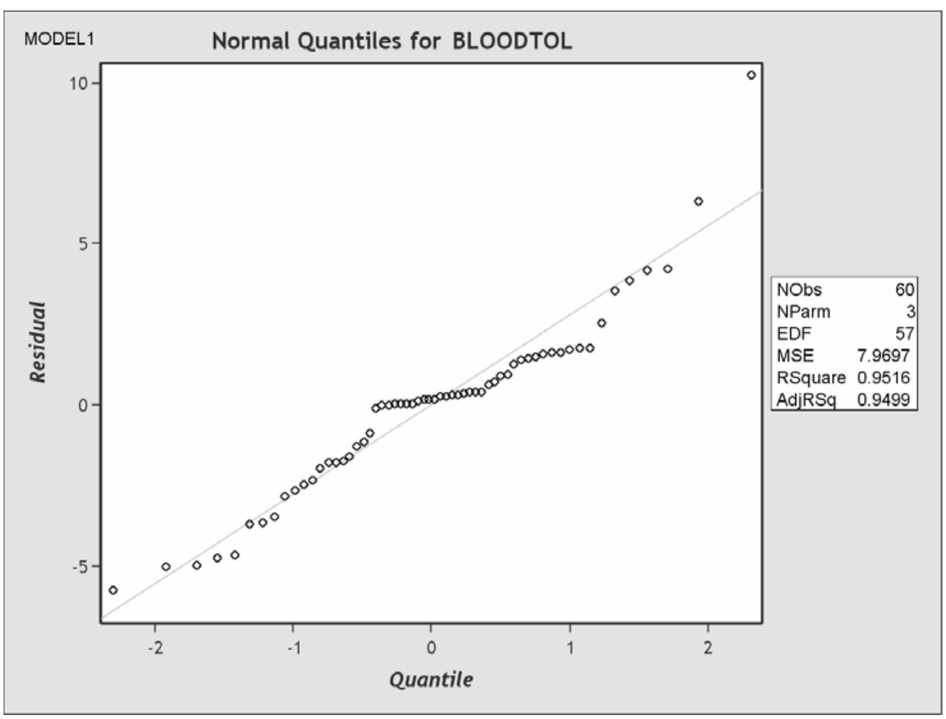
What is a normal QQ plot? (cont.)

- If data come from a normal distribution, then the quantiles of their standardized values should be approximately equivalent to the known quantiles of the standard normal distribution.
- A normal QQ plot graphs the quantiles of the data against the known quantiles of the standard normal distribution. Since we expect the quantiles to be roughly equivalent, then the QQ plot should follow the 45° reference line.

Normal QQ plot of residuals in SAS



What a normal QQ plot shouldn't look like ...



The Wilk-Shapiro test

H_0: The data are normally distributed
H_A: The data are not normally distributed

```
proc reg data = one noprint;
    model infrisk = los cult beds;
    output out = fitdata rstudent = jackknife;
run;
quit;

proc univariate data = fitdata normal;
    var jackknife;
run;
```


The Wilk-Shapiro test (cont.)

```
              Tests for Normality

Test          -Statistic---    -----p Value------

Shapiro-Wilk      0.994445    Pr < W       0.9347
```

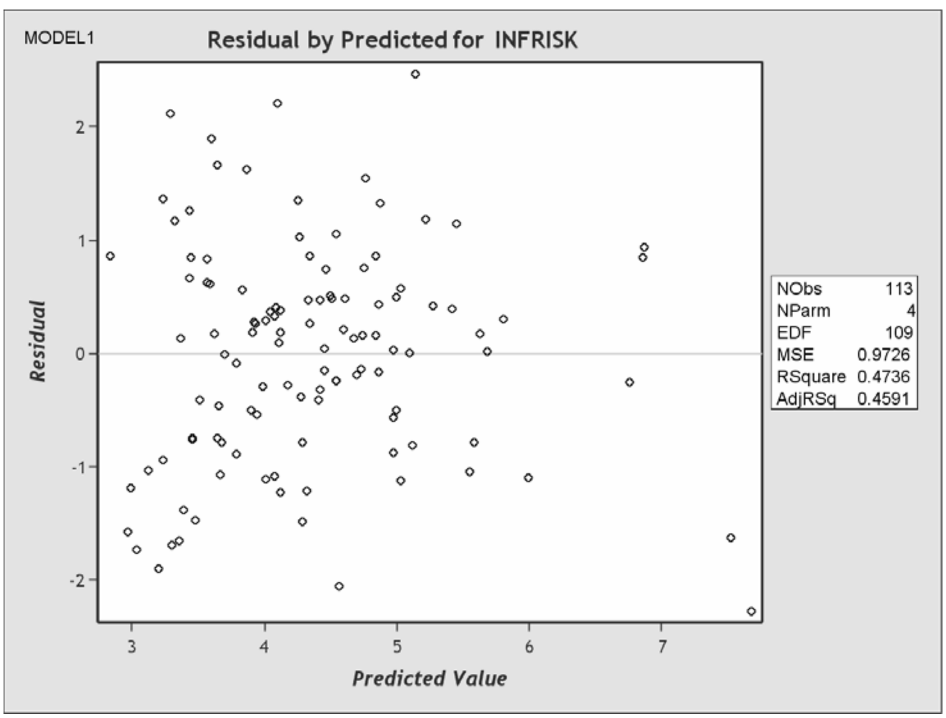
We fail to reject the null hypothesis and conclude that there is insufficient evidence to conclude that the model errors are not normally distributed.

Identifying departures from homoscedasticity

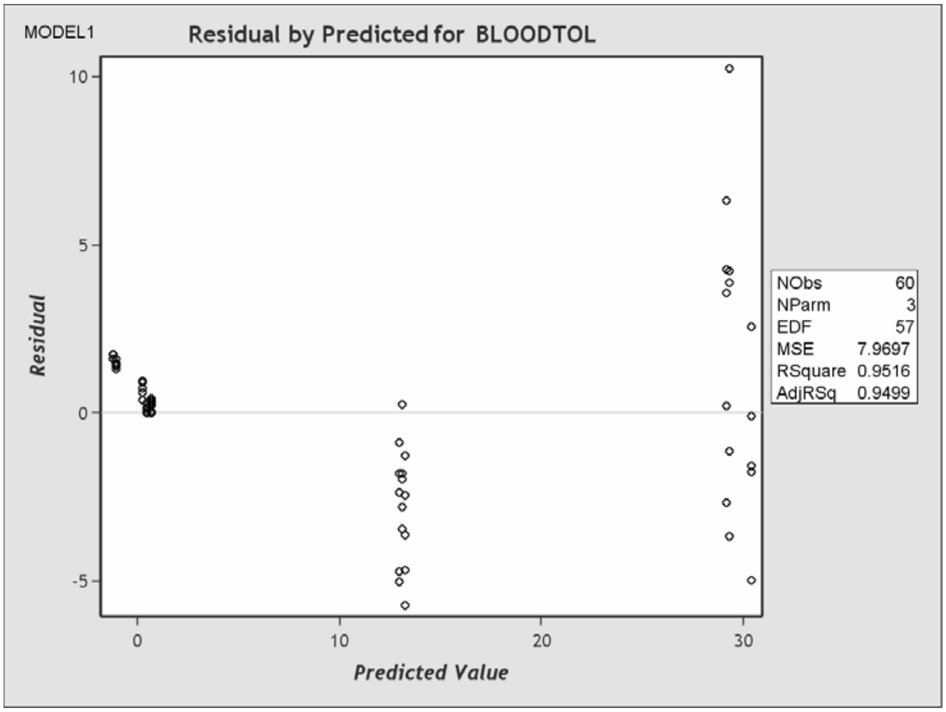
We identify departures from homoscedasticity by plotting the residuals versus the predicted values.

```
ods html style = Journal;
ods graphics on;
ods select ResidualByPredicted;

proc reg data = one plots(unpack);
    model infrisk = los cult beds;
run;
quit;

ods graphics off;
ods html close;
```


Departures from homoscedasticity (cont.)



The “megaphone” plot ... (i.e. a violation)

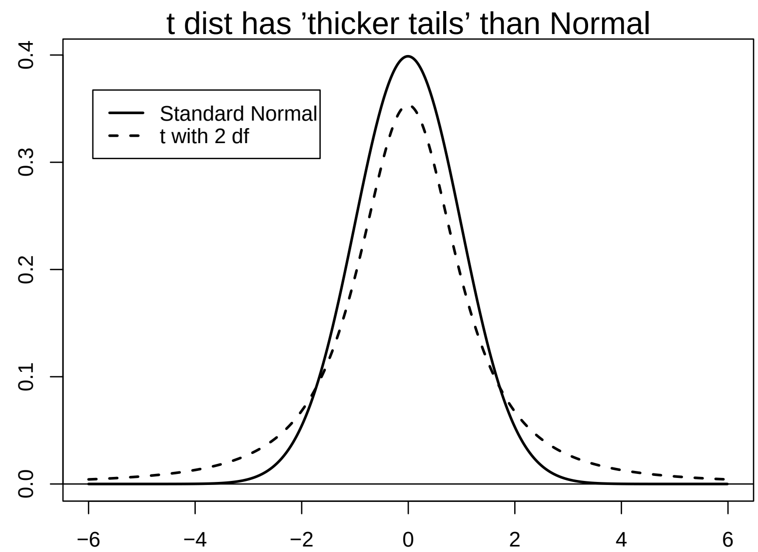


Outliers

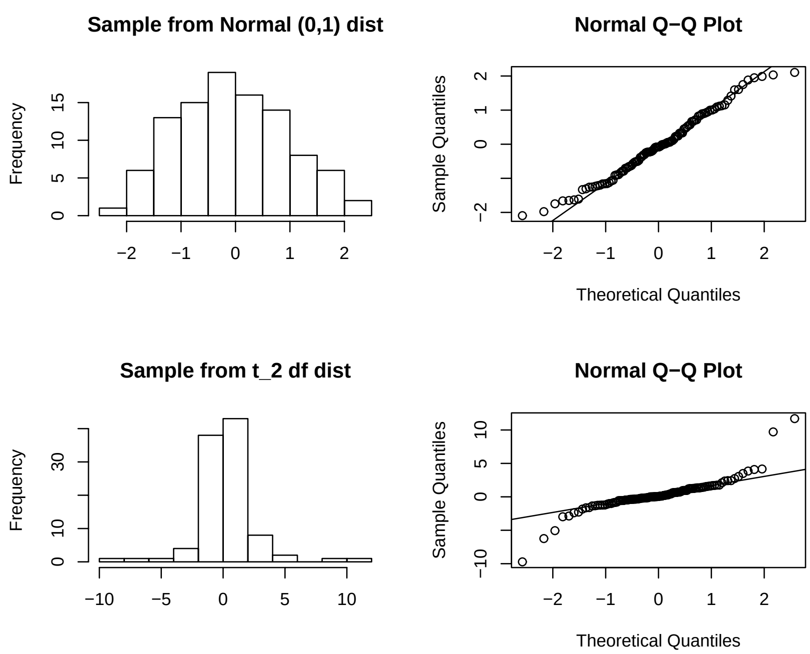
Outliers are observations that are *extreme* in the sense that they are noticeably different than the other data points. What causes outliers?

1. Data entry errors (the biggest culprit!)
2. Data do not represent a homogeneous set to which a single model applies. Rather, the data are a heterogeneous set of two or more types, of which one is more frequent than the others.
3. Error distributions that have “thick tails” in which extreme observations occur with greater frequency. (What does that mean?)

Outliers (cont.)



Outliers (cont.)



Outliers (cont.)

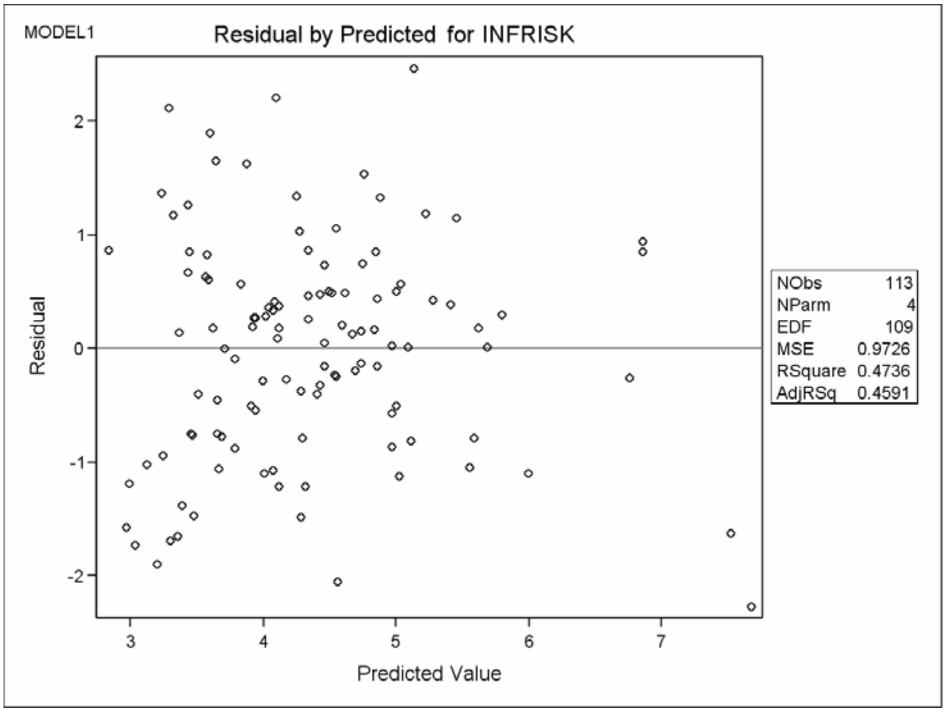
Although linear regression is robust to departures from normality, this is not the case when the error distribution has thick tails. Ironically, sampling distributions that look quite different from a normal distribution cause little trouble, while these thick tail distributions flaw the inference based on F tests.

Outlier detection - visual means

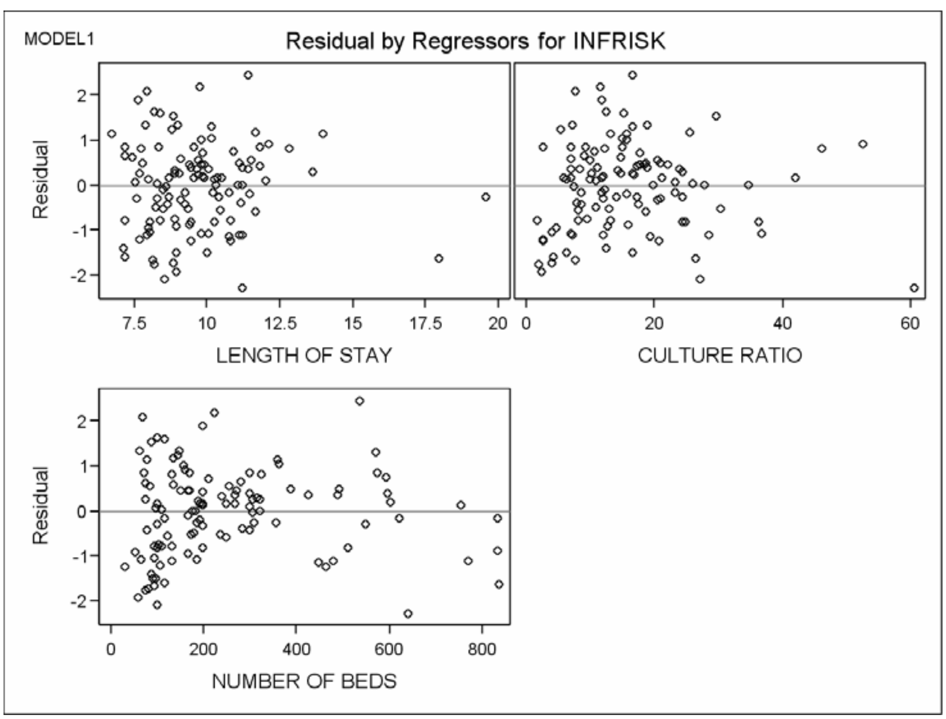
1. Simple scatterplots of the data (useful primarily for SLR)
2. Plots of the residuals versus the fitted values
3. Plot of the residuals versus each predictor

```
ods html style = analysis;
ods graphics on;
ods select ResidualByPredicted;
ods select ResidualPanel1;

proc reg data = one plots(unpack);
    model infrisk = los cult beds;
run; quit;

ods graphics off;
ods html close;
```

Outlier detection - Resid's vs. predicted values



Outlier detection - Resid's vs. predictors



Outlier detection - numerical means

Some rules of thumb about jackknife residuals

- Jackknife residuals with a magnitude less than 2 (i.e. between -2 and +2) are not unusual.
- Jackknife residuals with a magnitude greater than 2 deserve a look.
- Jackknife residuals with a magnitude greater than 4 are highly suspect.

Outlier detection in SAS

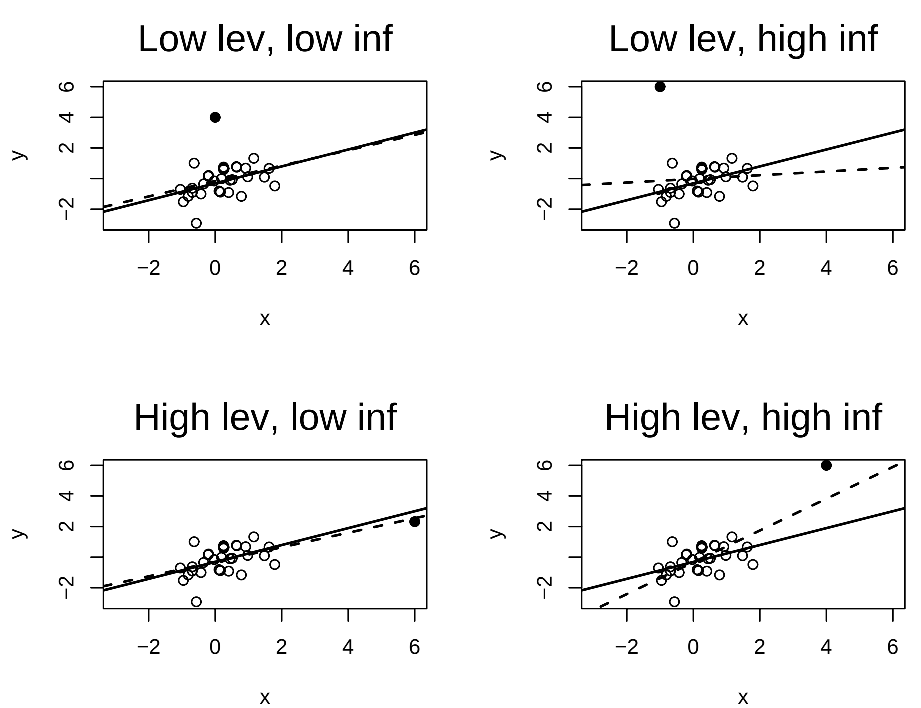
```
proc reg data = one;
    model infrisk = los cult beds;
    output out = fitdata rstudent = jackknife;
run;
quit;

proc print data = fitdata;
    where abs(jackknife) > 2;
run;
```

Obs	INFRISK	LOS	CULT	BEDS	jackknife
8	5.4	11.18	60.5	640	-2.66653
35	6.3	9.74	11.4	221	2.29488
53	7.6	11.41	16.6	535	2.60976
63	5.4	7.93	7.5	68	2.20954
96	2.5	8.54	27.0	98	-2.15224

Leverage and influence

- The *leverage* of a data point refers to how extreme it is relative to $\bar{x}$. Observations far away from $\bar{x}$ are said to have “high leverage”.
- The *influence* of a data point refers to its impact on the fitted regression line. If an observation “pulls” the regression line away from the fitted line that *would* have resulted if that point had not been present, then that observation is deemed “influential”.

Dotted = with point; Solid = without point



Measuring leverage

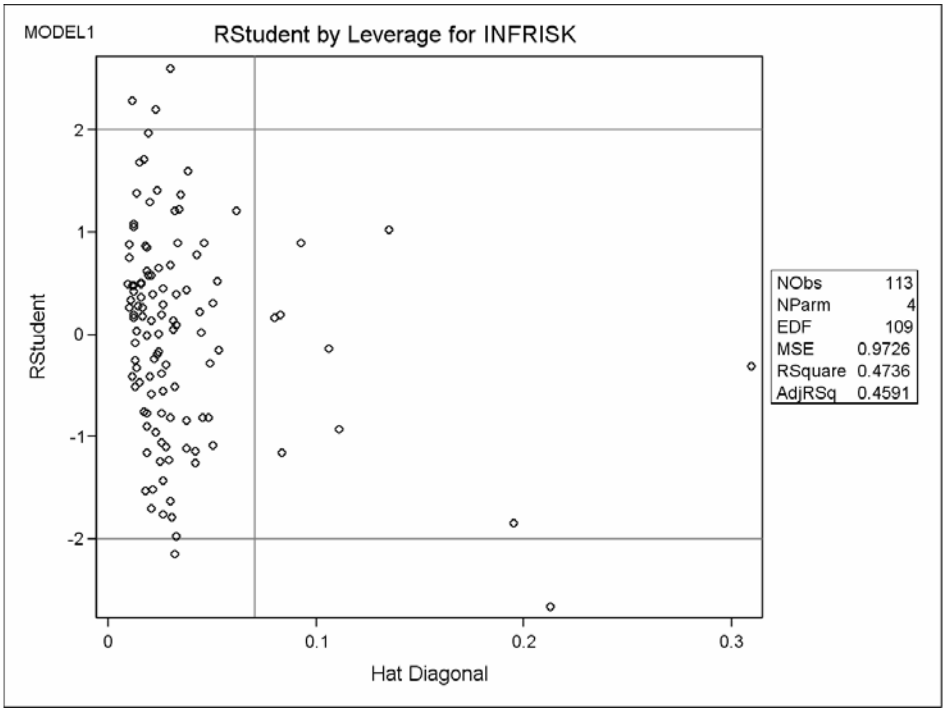
For each data point, there is a corresponding quantity known as the *hat diagonal* that measures the standardized distance of the ith observation to the center of the predictors. In symbols, the hat diagonal is written h_i. (We saw this term in the definition of studentized residuals.)

As a general rule of thumb, any value of h_i greater than $2(k+1)/n$ is cause for concern, where k is the number of predictors in the model and n is the number of data points.

Leverage and residual plots in SAS

```
ods html style = analysis;
ods graphics on;
ods select RStudentByLeverage;

proc reg data = one plots(unpack);
    model infrisk = los cult beds;
run;
quit;

ods graphics off;
ods html close;
```

Leverage and residual plots in SAS



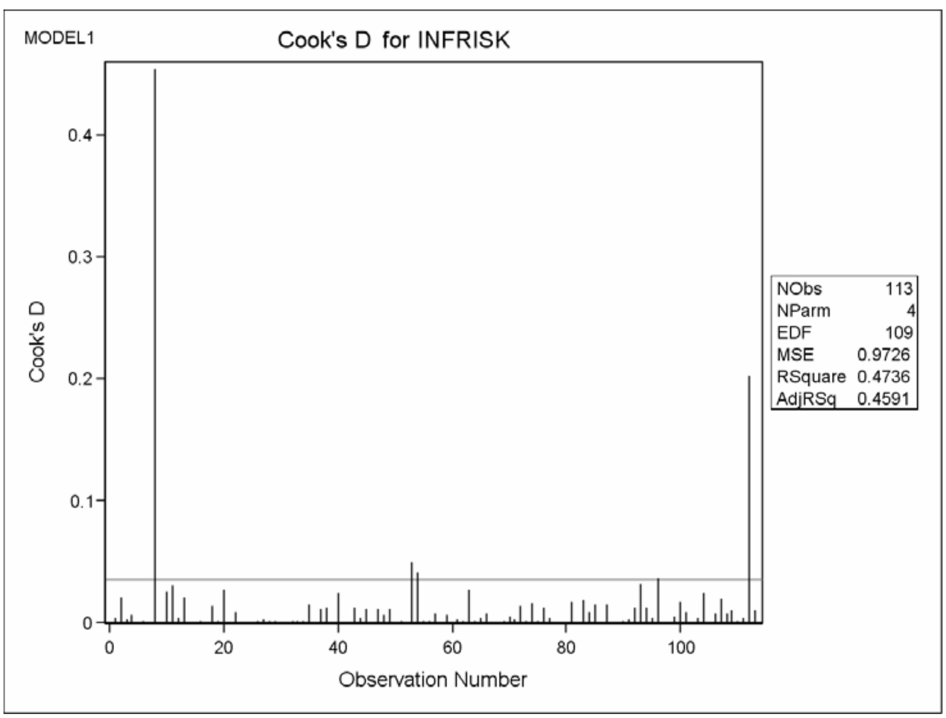
Measuring influence

For each data point, there is a corresponding quantity known as *Cook's D*$_i$ ('D' for 'distance') that is a standardized distance measuring how far $\hat{\boldsymbol{\beta}}$ moves when the ith observation is removed. As a general rule of thumb, any observation with a value of Cook's D$_i$ greater than $4/n$, where n is the number of observations, deserves a closer look.

Graphing of Cooks D in SAS

```
ods html style = analysis;
ods graphics on;
ods select CooksD;

proc reg data = one plots(unpack);
    model infrisk = los cult beds;
run;
quit;

ods graphics off;
ods html close;
```


Graphing of Cooks D in SAS (cont.)



Leverage/influence - numerical assessment

```
proc reg data = one;
    model infrisk = los cult beds;
    output out = fitdata cookd = cooksd h = hat;
run;
quit;

*(2*4)/113 = 0.071;
proc print data = fitdata;
    where hat ge (2*4)/113;
run;

*4/113 = 0.035;
proc print data = fitdata;
    where cooksd ge 4/113;
run;
```

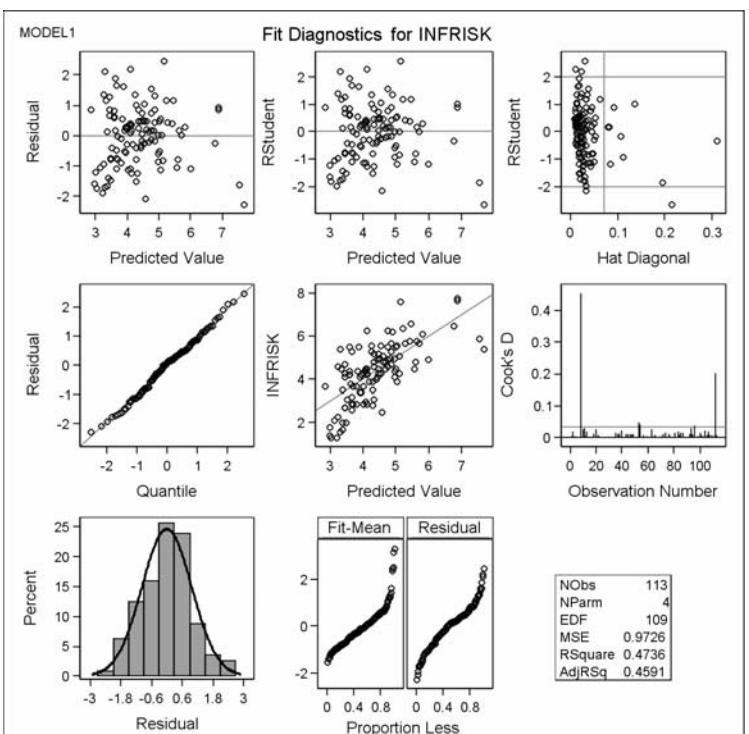

Leverage/influence - numerical assessment

Obs	INFRISK	LOS	CULT	BEDS	cooksd	hat	jackknife
8	5.4	11.18	60.5	640	0.45427	0.21252	-2.66653
11	4.9	11.07	28.5	768	0.03058	0.08368	-1.15910
13	7.7	12.78	46.0	322	0.02080	0.09230	0.90371
20	4.1	9.35	15.9	833	0.02697	0.11055	-0.93115
46	4.6	10.16	8.4	831	0.00055	0.10609	-0.13510
47	6.5	19.56	17.2	306	0.01081	0.30916	-0.30957
54	7.8	12.07	52.4	157	0.04106	0.13462	1.02773
78	4.9	10.23	9.9	752	0.00066	0.07977	0.17330
110	5.8	9.50	42.0	98	0.00083	0.08236	0.19197
112	5.9	17.94	26.4	835	0.20239	0.19506	-1.84793

All the graphics in one panel

```
ods html style = analysis;
ods graphics on;

proc reg data = one plots;
    model infrisk = los cult beds;
run;
quit;

ods graphics off;
ods html close;
```


All the graphics in one panel



Collinearity

Collinearity is a problem that exists when some (or all) of the independent variables are strongly linearly associated with one another. If collinearity exists in your data, then the following problems result.

- The estimated regression coefficients can be highly inaccurate.
- The standard errors of the coefficients can be highly inflated.
- The p-values, and all subsequent inference, can be wrong.

Symptoms of collinearity

- Large changes in coefficient estimates and/or in their standard errors when independent variables are added/deleted.
- Large standard errors.
- Non-significant results for independent variables that should be significant.
- Wrong signs on slope estimates.
- Overall test significant, but partial tests insignificant.
- Strong correlations between independent variables.
- Large *variance inflation factors* (VIFs).

Variance inflation factor

For each independent variable X_j in a model, the *variance inflation factor* is calculated as

$$\text{VIF}_j = \frac{1}{1 - R^2_{X_j \sim \text{all other } X\text{s}}}$$

where $R^2_{X_j \sim \text{all other } X\text{s}}$ is the usual R^2 obtained by regressing X_j on all the other Xs, and represents the proportion of variation in X_j that is explained by the remaining independent variables. Therefore, $1 - R^2_{X_j \sim \text{all other } X\text{s}}$ is a measure of the variability in X_j that *isn't* explained by the other independent variables.

Variance inflation factor (cont.)

When there is strong collinearity,

- $R^2_{X_j \sim \text{all other } X\text{s}}$ will be large
- $1 - R^2_{X_j \sim \text{all other } X\text{s}}$ will be small, and so
- VIF_j will be large.

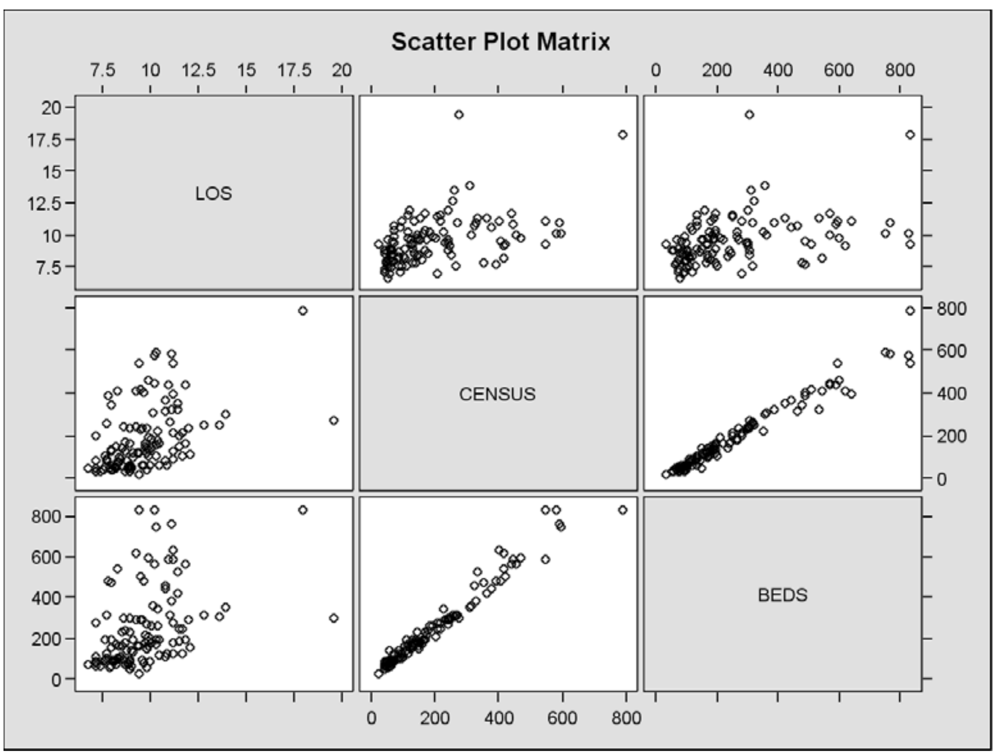
As a general rule of thumb, strong collinearity is present when $\text{VIF}_j > 10$.

Collinearity example

Consider the MLR of INFRISK on LOS, CENSUS and BEDS.

```
ods html;
ods graphics on;

proc corr data = one plots=matrix;
    var los census beds;
run;

ods graphics off;
ods html close;

proc reg data = one;
    model infrisk = los/vif;
    model infrisk = los census/vif;
    model infrisk = los census beds/vif;
run;
```


Collinearity: PROC CORR output



Collinearity: PROC CORR output (cont.)

Pearson Correlation Coefficients, N = 113
Prob > |r| under H0: Rho=0

	LOS	CENSUS	BEDS
LOS	1.00000	0.47389	0.40927
LENGTH OF STAY		<.0001	<.0001
CENSUS	0.47389	1.00000	0.98100
AVG DAILY CENSUS	<.0001		<.0001
BEDS	0.40927	0.98100	1.00000
NUMBER OF BEDS	<.0001	<.0001	

Collinearity: PROC REG output

INFRISK REGRESSED ON LOS

Root MSE	1.13929	R-Square	0.2846
Dependent Mean	4.35487	Adj R-Sq	0.2781
Coeff Var	26.16119		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.74430	0.55386	1.34	0.1817	0
LOS	1	0.37422	0.05632	6.64	<.0001	1.00000

Collinearity: PROC REG output (cont.)

INFRISK REGRESSED ON LOS AND CENSUS

Root MSE	1.12726	R-Square	0.3059
Dependent Mean	4.35487	Adj R-Sq	0.2933
Coeff Var	25.88504		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.99950	0.56531	1.77	0.0798	0
LOS	1	0.31908	0.06328	5.04	<.0001	1.28960
CENSUS	1	0.00145	0.00078668	1.84	0.0687	1.28960

Collinearity: PROC REG output (cont.)

INFRISK REGRESSED ON LOS, CENSUS AND BEDS

Root MSE	1.12953	R-Square	0.3094
Dependent Mean	4.35487	Adj R-Sq	0.2904
Coeff Var	25.93727		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.82296	0.61382	1.34	0.1828	0
LOS	1	0.33538	0.06706	5.00	<.0001	1.44245
CENSUS	1	-0.00142	0.00392	-0.36	0.7177	31.90045
BEDS	1	0.00225	0.00302	0.75	0.4569	29.71362