Lecture 3: Inference and Diagnostics for Simple Linear Regression

Nancy R. Zhang

Statistics 203, Stanford University

January 12, 2010

Review

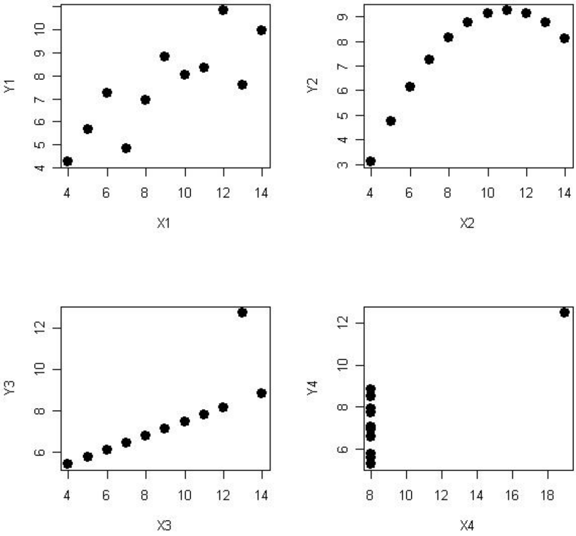
Assumptions of the linear model

$$Y_i = \beta_0 X_i + \beta_1 + \epsilon_i$$

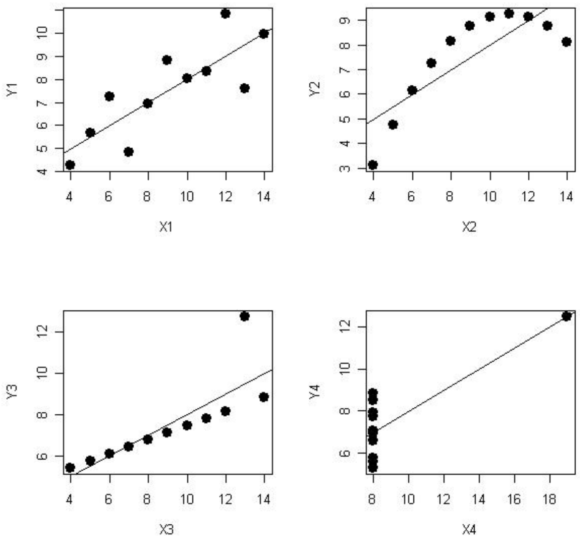
1. Y has a linear dependence on X.
2. Error variances are equal.
3. Errors are independent.
4. Errors are Gaussian.

Data points that deviate from the “bulk", which we assume to satisfy these model assumptions, are called outliers.

Anscombe's quartet



Anscombe's quartet



Standardized Residuals

A simple way to evaluate model fit is through the residuals,

$$r_i = y_i - \hat{y}_i.$$

The residuals have variance

$$\sigma^2(1 - h_{ii}),$$

where

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}$$

is called the *leverage*.

We need to standardize the residuals so that they are comparable:

$$z_i = \frac{r_i}{\sigma\sqrt{1 - h_{ii}}}.$$

However, we don't know σ, so need to estimated it.

Standardized Residuals

There are two ways to estimate σ:

1. Using all of the data:

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

2. Let $SSE_{(i)}$ be the sum of squared residuals obtained when we fit the model to the sample with the i-th point taken out.

$$\hat{\sigma}^2_{(i)} = \frac{SSE_{(i)}}{n-3},$$

Thus, there are two ways to standardize r_i:

1. Studentized residuals: $r_i = \frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$
2. Studentized deleted residuals:

$$r_i^* = \frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} \qquad \text{has } t_{n-2} \text{ distribution.}$$

New York Rivers Data

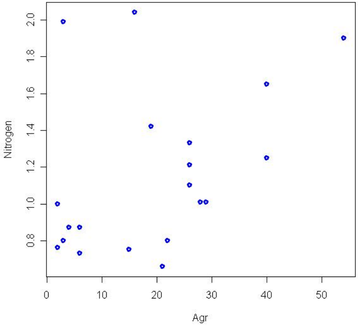
Haith (1976) Study: How does land use around a river basin contribute to water pollution?

River	Forest	CommIndl	Nitrogen
Olean	63	0.29	1.1
Cassadaga	57	0.09	1.01
Oatka	26	0.58	1.9
Neversink	84	1.98	1
Hackensack	27	3.11	1.99
Wappinger	61	0.56	1.42
Fishkill	60	1.11	2.04
Honeoye	43	0.24	1.65
Susquehanna	62	0.15	1.01
Chenango	60	0.23	1.21
Tioughnioga	53	0.18	1.33
WestCanada	75	0.16	0.75
EastCanada	84	0.12	0.73
Saranac	81	0.35	0.8
Ausable	89	0.35	0.76
Black	82	0.15	0.87
Schoharie	70	0.22	0.8
Raquette	75	0.18	0.87
Oswegatchie	56	0.13	0.66
Cohocton	49	0.13	1.25

CommIndl: % land area in either commercial or industrial use

Nitrogen: Mean nitrogen concentration (mg/liter)

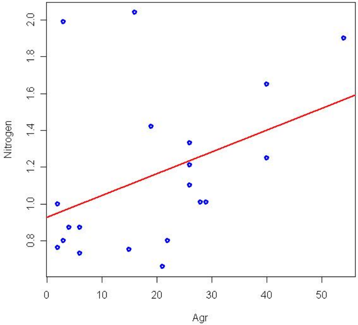
Diagnostic plots - scatter plot



Fit model:

$$\text{Nitrogen} = \beta_0 + \beta_1 \text{Agr} + \text{error}$$

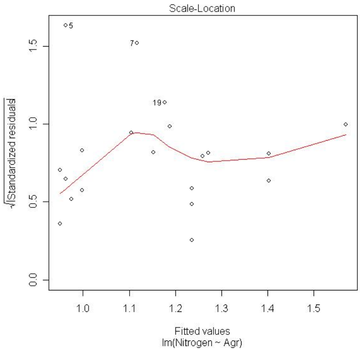
Diagnostic plots - scatter plot



Fit model:

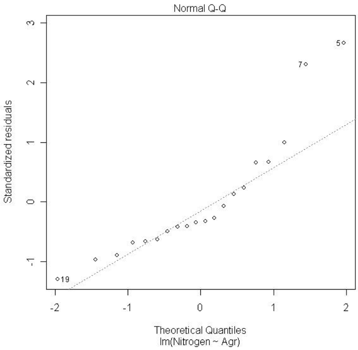
$$\text{Nitrogen} = \beta_0 + \beta_1 \text{Forest} + \text{error}$$

Diagnostic plots - standardized residual plot



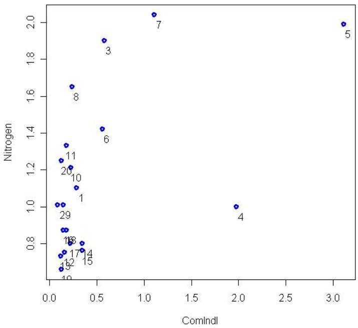
Plot standardized residuals versus X_i, large standardized residuals indicate outliers.

Diagnostic plots - qq-plot of residuals



Quantile-quantile plots may be sometimes helpful for identifying outliers.

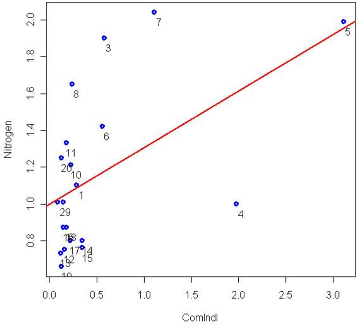
Diagnostic plots - scatter plot



Fit model:

$$\text{Nitrogen} = \beta_0 + \beta_1 \text{CommInd} + \text{error}$$

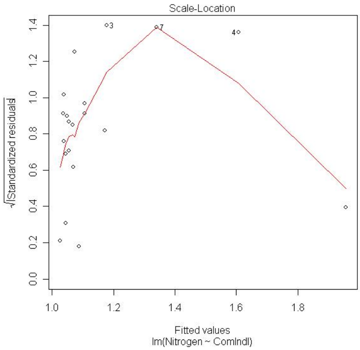
Diagnostic plots - scatter plot



Fit model:

$$\text{Nitrogen} = \beta_0 + \beta_1 \text{CommInd} + \text{error}$$

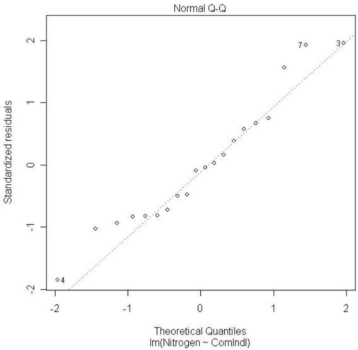
Diagnostic plots - scatter plot



Fit model:

$$\text{Nitrogen} = \beta_0 + \beta_1 \text{CommInd} + \text{error}$$

Diagnostic plots - scatter plot



Fit model:

$$\text{Nitrogen} = \beta_0 + \beta_1 \text{CommInd} + \text{error}$$

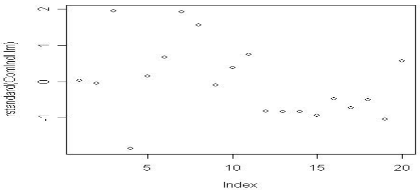
Outliers in X versus Outliers in Y

1. Outliers in Y can be detected by examining the standardized residuals.
$$r_i = \frac{r_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \qquad \text{easier to compute}$$
2. Outliers in X can be detected using leverage values.
$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

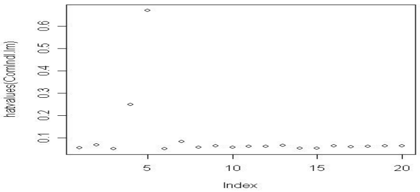
   Mean of the leverage values is $2/n$. When data set is reasonably large, leverage values larger than $2\times$(mean leverage) are considered large. Another convention is to consider leverage between 0.2 and 0.5 as high.

Leverage versus Residuals

Standardized Residuals

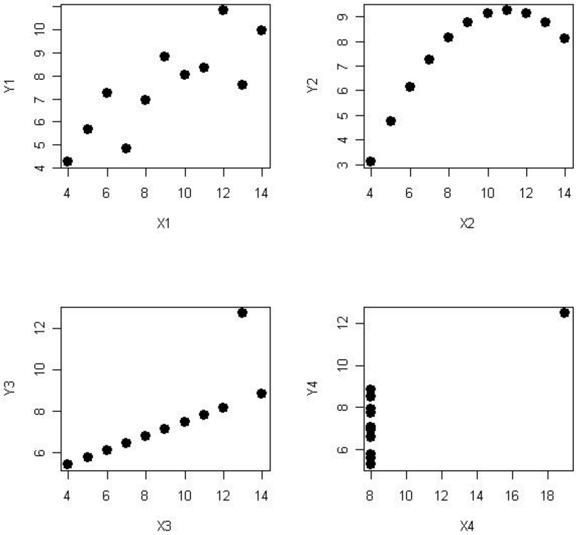


Leverage



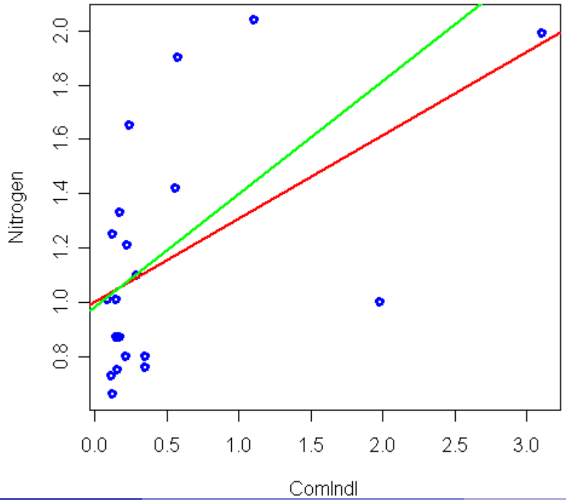
Is there a measure that can detect both types of outliers?

Anscombe's quartet



Influence of Outliers

How much influence does the data point have on the model fit?



Different measures of Influence

1. How much influence does observation i have on its own fit?

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

   DFFITS exceeding $2\sqrt{p/n}$ is considered large.

2. How much influence does observation i have on the fitted β's?

$$(DFBETAS)_i = \frac{\hat{\beta}_1 - \hat{\beta}_{1(i)}}{\sqrt{MSE_{(i)} c_{11}^{-1}}},$$

   where $c_{11} = \sum_i (x_i - \bar{x})^2$. *DFBETA* exceeding $2/\sqrt{n}$ is considered large.

These conventional rules work for "reasonably sized" data sets.

Different measures of Influence: Cook's Distance

1. Cook's distance is defined as:

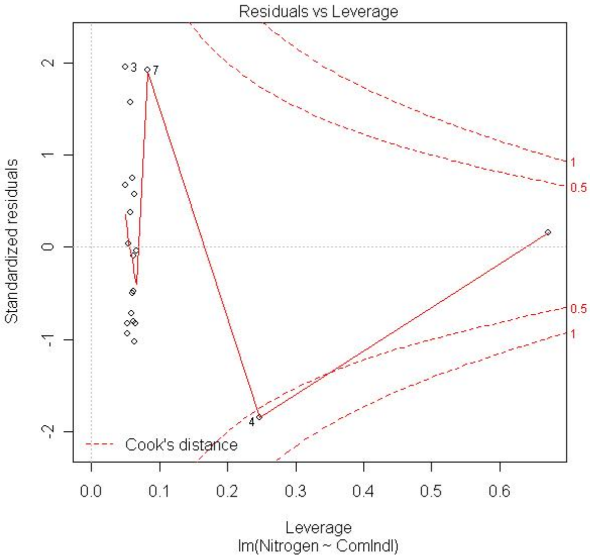
$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{pMSE}$$

2. Considers the influence of Y_i on all of the fitted values, not just the i-th case.
3. It can be shown that D_i is equivalent to

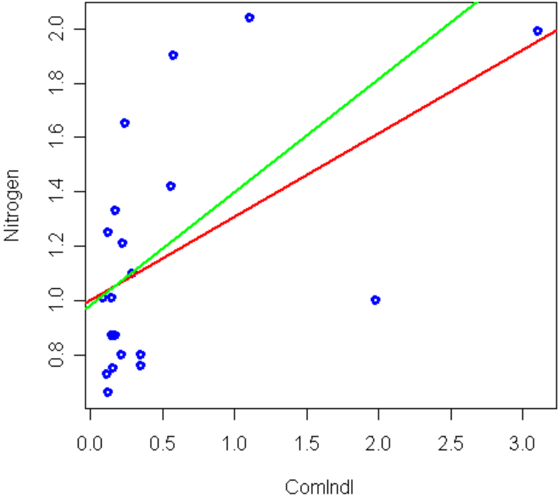
$$\frac{r_i^2}{pMSE} \frac{h_{ii}}{(1 - h_{ii})^2}.$$

4. Compare D_i to the $F_{p,n-p}$ distribution.

Cook's Distance



Masking of Outliers



What to do with outliers?

1. Sometimes they hint that our model assumptions are wrong.
2. Down-weight or delete outlying data points.
3. Transform the variables (more on this later).