Statistics 191: Introduction to Applied Statistics

Diagnostics & Influence

Jonathan Taylor
Department of Statistics
Stanford University

February 22, 2010

Diagnostics in multiple linear model

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University

Outline

- Diagnostics – again
- Different residuals
- Influence
- Outlier detection
- Residual plots:
  - partial regression (added variable) plot,
  - partial residual (residual plus component) plot

Scottish hill races data

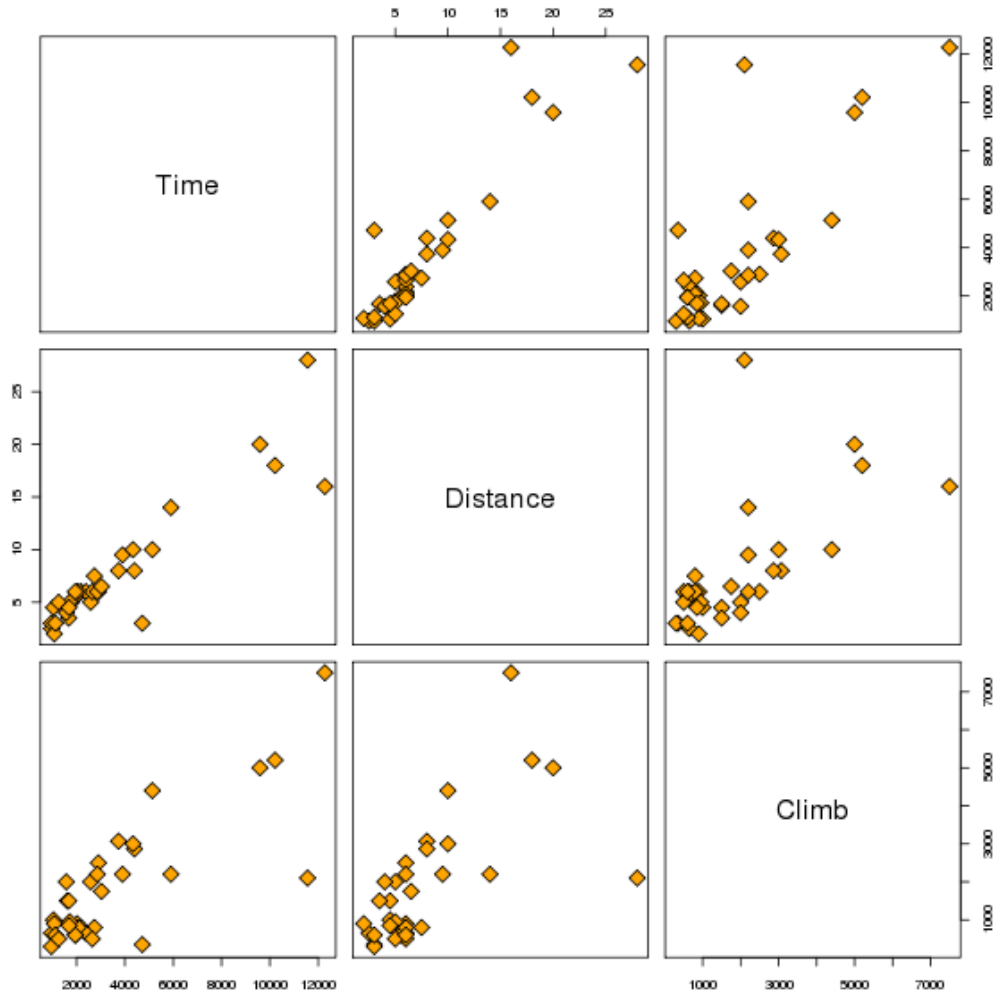
Description

Variable	Description
`Time`	Record time to complete course
`Distance`	Distance in the course
`Climb`	Vertical climb in the course

Scottish hill races data

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Diagnostics

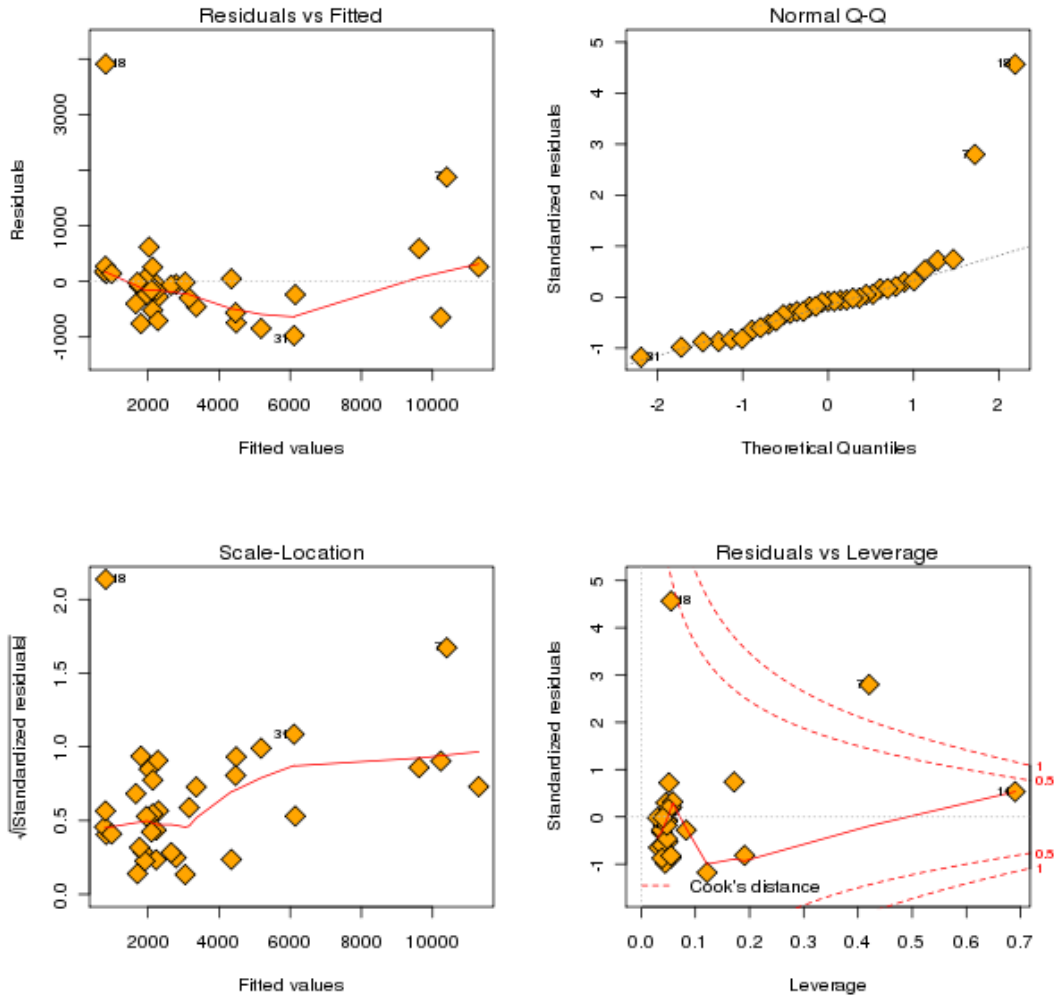
What can go wrong?

- Regression function can be wrong: maybe regression function should be quadratic (see R code).
- Model for the errors may be incorrect:
  - may not be normally distributed.
  - may not be independent.
  - may not have the same variance.
- Detecting problems is more *art* then *science*, i.e. we cannot *test* for all possible problems in a regression model.
- Basic idea of diagnostic measures: if model is correct then residuals $e_i = Y_i - \widehat{Y}_i, 1 \leq i \leq n$ should look like a sample of (not quite independent) $N(0, \sigma^2)$ random variables.

Standard diagnostic plots

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Problems with the errors

Possible problems & diagnostic checks

- Errors may not be normally distributed or may not have the same variance – `qqnorm` can help with this. This may not be too important in large samples.
- Variance may not be constant. Can also be addressed in a plot of $\boldsymbol{X}$ vs. $\boldsymbol{e}$: *fan shape* or other trend indicate non-constant variance.
- Influential observations. Which points "affect" the regression line the most?
- Outliers: points where the model really does not fit! Possibly mistakes in data transcription, lab errors, who knows? Should be recognized and (hopefully) explained.

Residuals

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University

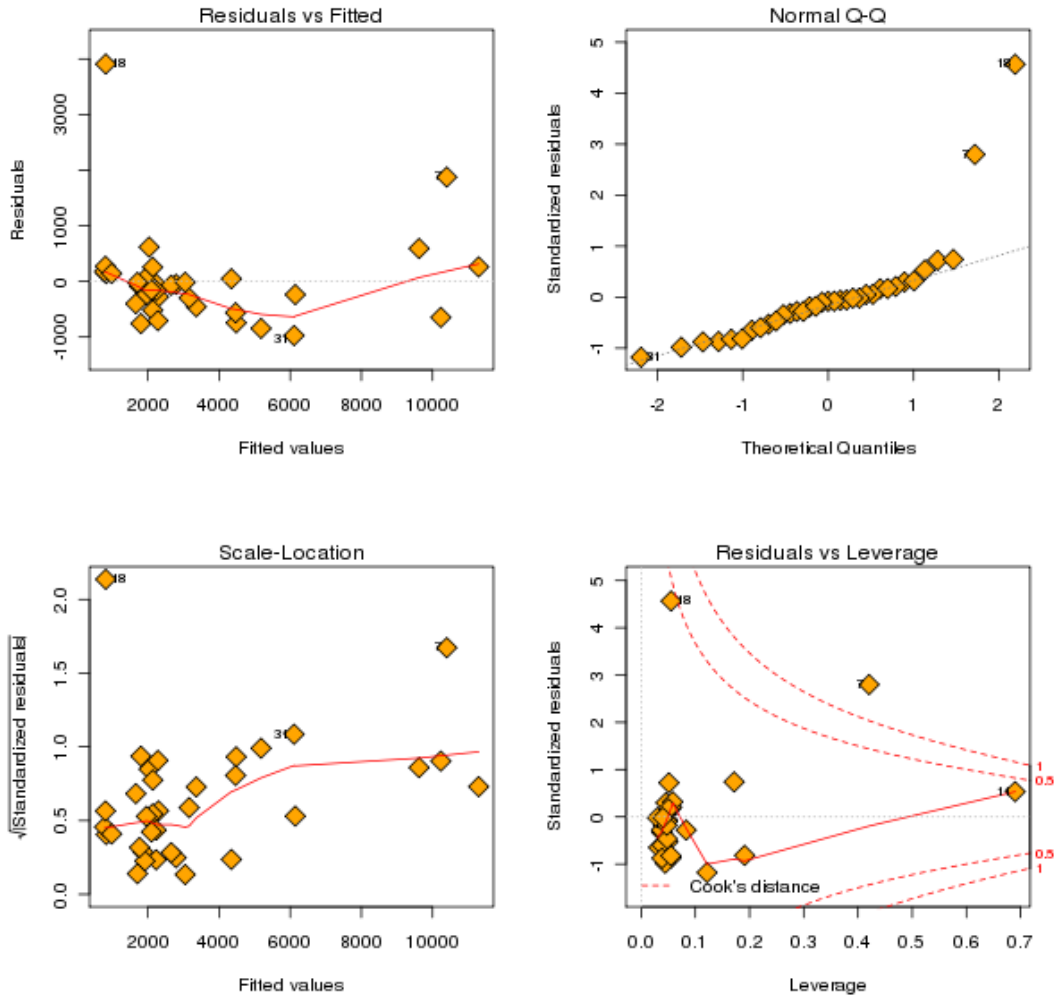
Types of residuals

- Ordinary residuals: $e_i = Y_i - \widehat{Y}_i$. These measure the deviation of predicted value from observed value, but their distribution depends on unknown scale, σ.
- Internally studentized residuals (`rstandard` in `R`): $r_i = e_i / s(e_i) = e_i / \widehat{\sigma}\sqrt{1 - H_{ii}}$, H is the "hat" matrix. These are almost t-distributed, except $\widehat{\sigma}$ depends on e_i.
- Externally studentized residuals (`rstudent` in `R`): $t_i = e_i / \widehat{\sigma_{(i)}}\sqrt{1 - H_{ii}} \sim t_{n-p-2}$. These are exactly t distributed so we know their distribution and can use them for tests, if desired.

Standard diagnostic plots

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Influence of an observation

Dropping an observation

- In this setting, a $\cdot_{(i)}$ indicates i-th observation was not used in fitting the model.
- For example: $\widehat{Y}_{j(i)}$ is the regression function evaluated at the j-th observations predictors BUT the coefficients $(\widehat{\beta}_{0(i)}, \ldots, \widehat{\beta}_{p(i)})$ were fit after deleting i-th row of data.
- Idea: if $\widehat{Y}_{j(i)}$ is very different than $\widehat{Y}_j$ (using all the data) then i is an influential point, at least for estimating the regression function at $(X_{1,j}, \ldots, X_{p,j})$.

Influence of an observation

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University

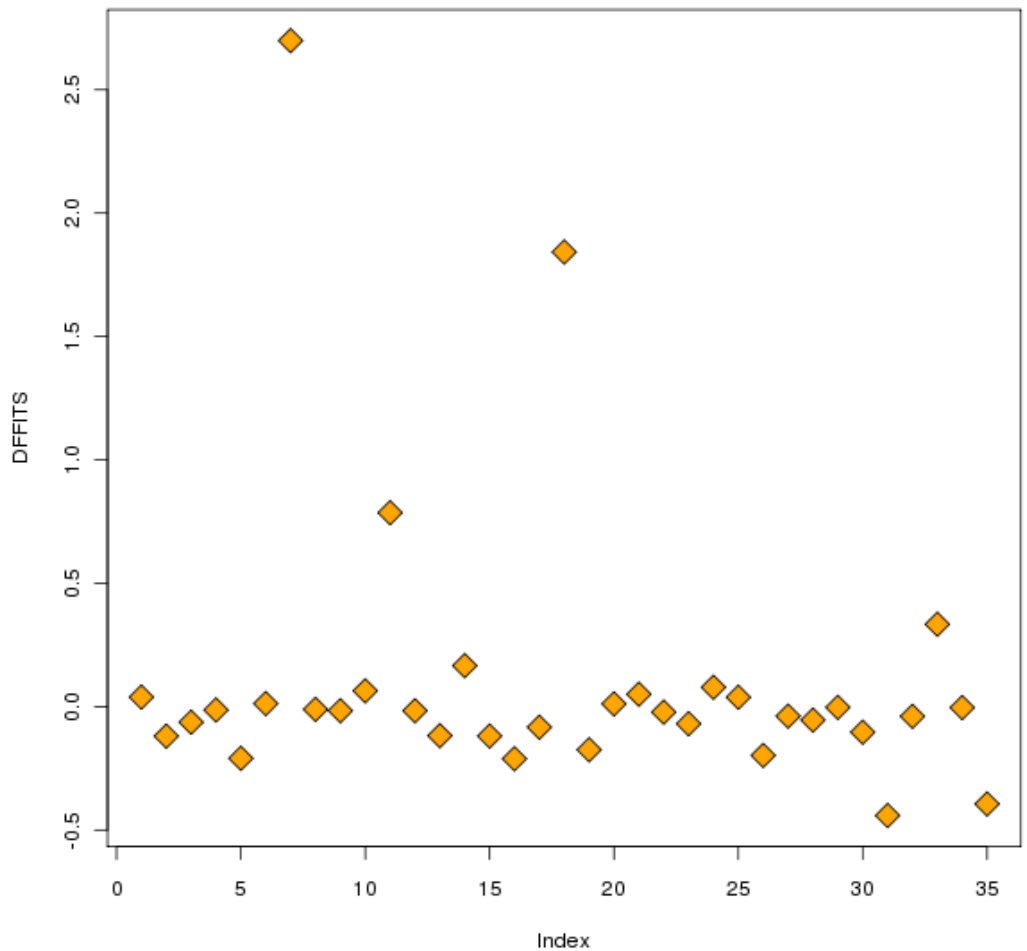
DFFITS

- $$DFFITS_i = \frac{\widehat{Y}_i - \widehat{Y}_{i(i)}}{\widehat{\sigma}_{(i)}\sqrt{H_{ii}}}$$
- This quantity measures how much the regression function changes at the i-th observation when the i-th variable is deleted.
- For small/medium datasets: value of 1 or greater is "suspicious". For large dataset: value of $2\sqrt{(p+1)/n}$.

Influence of an observation: DFFITS

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Influence of an observation

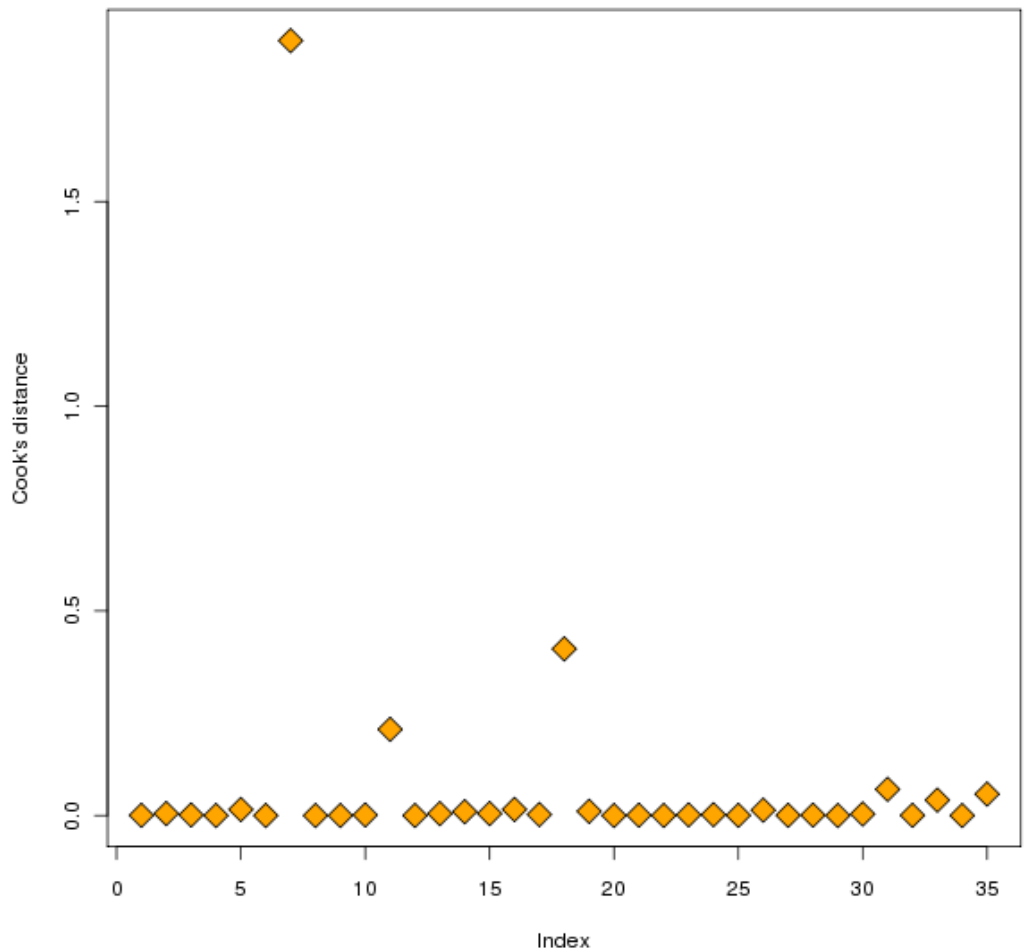
Cook's distance

- $$D_i = \frac{\sum_{j=1}^{n}(\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{(p+1)\,\widehat{\sigma}^2}$$
- This quantity measures how much the entire regression function changes when the i-th variable is deleted.
- Should be comparable to $F_{p+1,n-p-1}$: if the "p-value" of D_i is 50 percent or more, then the i-th point is likely influential: investigate further.

Influence of an observation: Cook's distance

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Influence of an observation

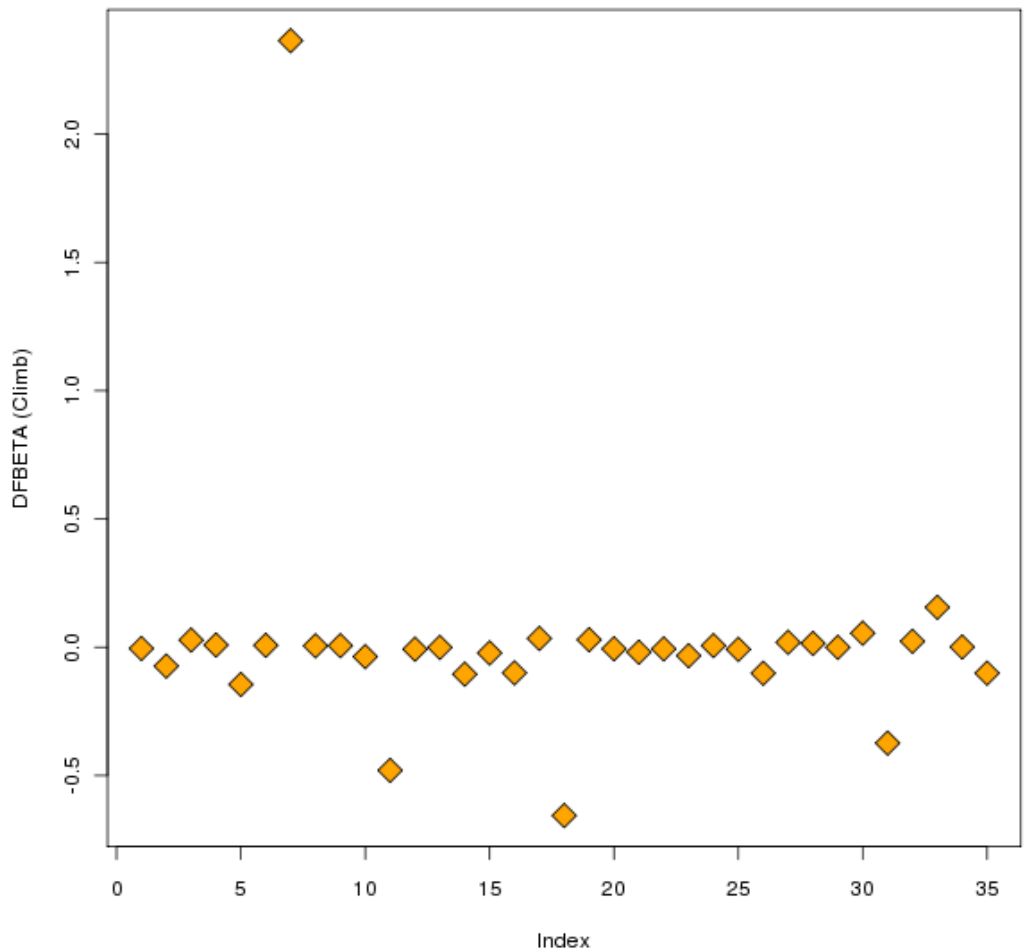
DFBETAS

- $$DFBETAS_{j(i)} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j(i)}}{\sqrt{\widehat{\sigma}^2_{(i)}(X^T X)^{-1}_{jj}}}.$$
- This quantity measures how much the coefficients change when the i-th variable is deleted.
- For small/medium datasets: value of 1 or greater is "suspicious". For large dataset: value of $2/\sqrt{n}$.

Influence of an observation: DFBETA, Climb

Statistics 191: Introduction to Applied Statistics

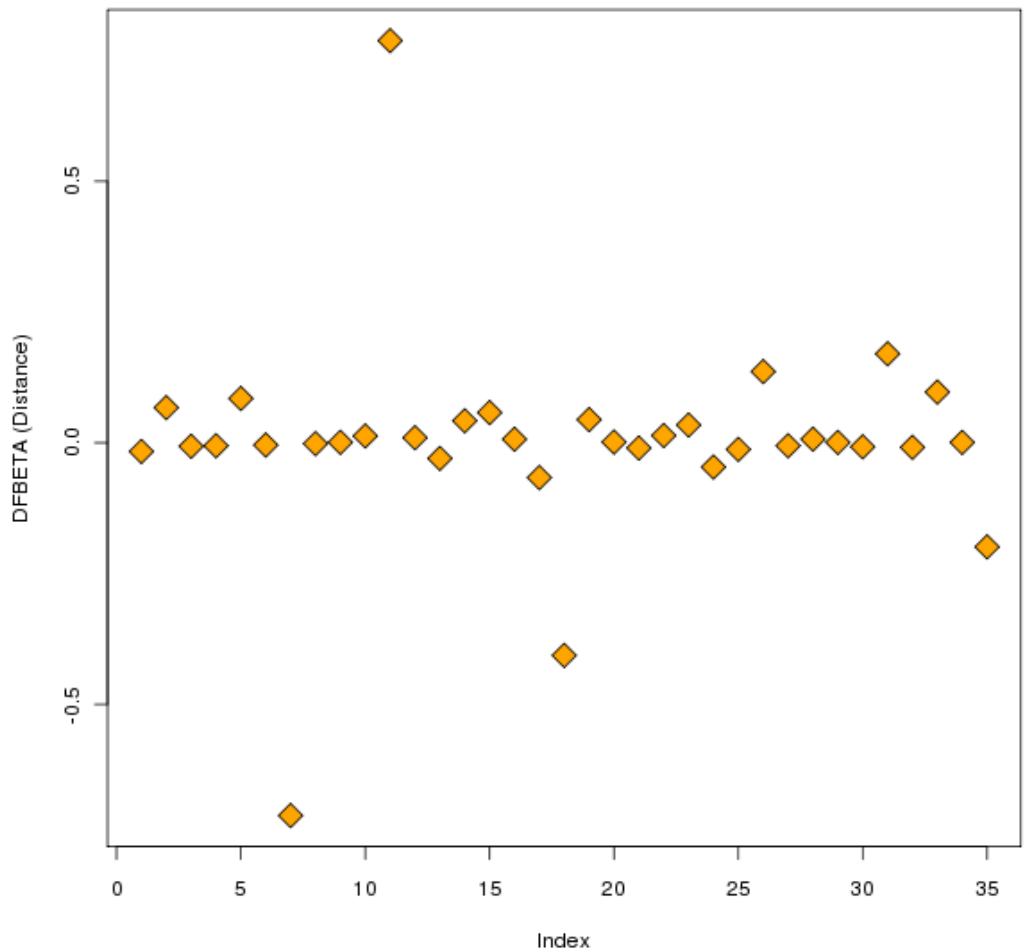
Jonathan Taylor
Department of Statistics
Stanford University



Influence of an observation: DFBETA, Distance

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Outliers

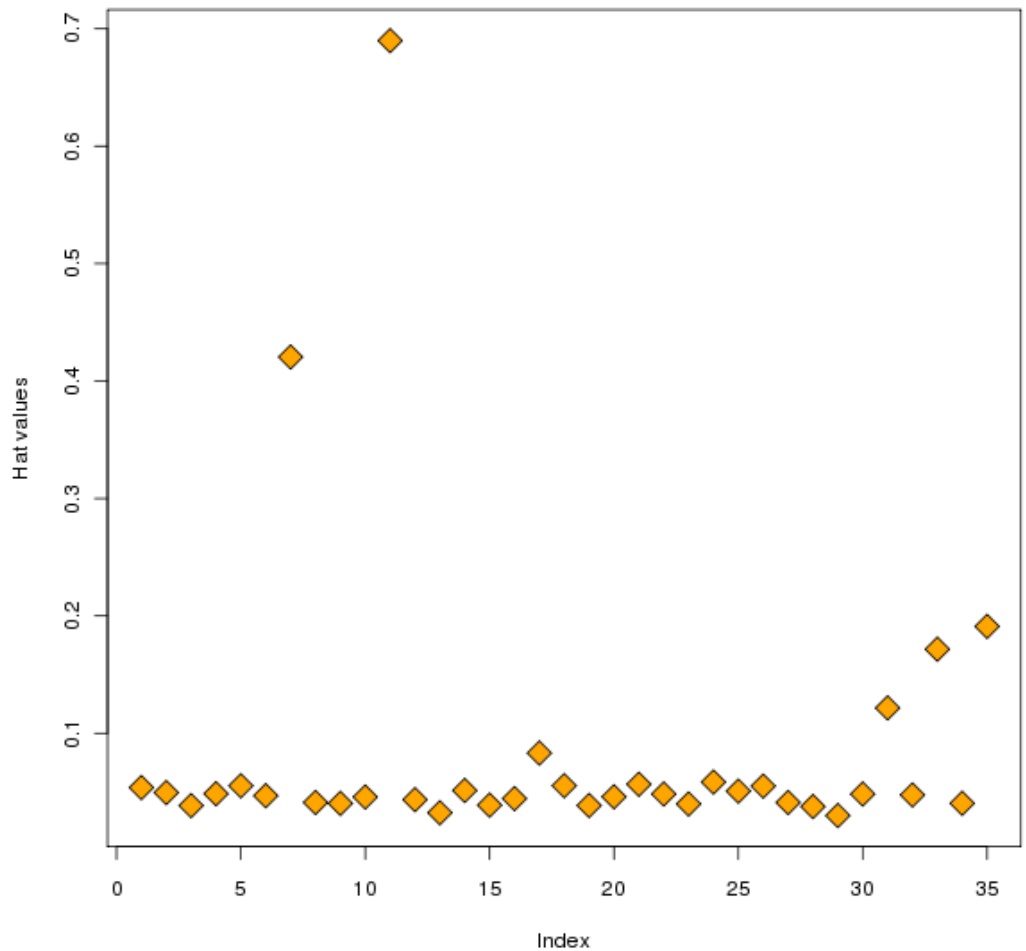
Basic definition

- *Outlier:* an observation pair $(Y, X_1, \ldots, X_p)$ that does not follow the model, while most other observations seem to follow the model.
- Outlier in predictors: the X values of the observation may lie outside the "cloud" of other X values. This means you may be extrapolating your model inappropriately. The values H_{ii} can be used to measure how "outlying" the X values are.
- Outlier in response: the Y value of the observation may lie very far from the fitted model. If the studentized residuals are large: observation may be an outlier.

Outlying X values

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Outliers

Crude (response) outlier detection test

- Strategy to detect outliers: "flag" large residuals.
- Problem: if n is large, if we "threshold" at $t_{1-\alpha/2,n-p-2}$ we will get many outliers by chance even if model is correct. In fact, we expect to see $n \cdot \alpha$ "outliers" by this test. Every large data set would have outliers in it, even if model was entirely correct!
- Problem is known as *multiple comparisons* or *simultaneous inference.* We are performing n hypothesis tests, but would still like to control the probability of making *any* false positive errors.
- One solution: Bonferroni correction, threshold at $t_{1-\alpha/(2*n),n-p-2}$.

Multiple comparisons

Bonferroni correction

- If we are doing many t (or other) tests, say $m >> 1$ we can control overall false positive rate at α by testing each one at level α/m.
- Proof, when the model is correct, with studentized residuals T_i:

$$
\begin{aligned}
& P(\text{at least one false positive}) \\
&= P\left(\cup_{i=1}^{m} |T_i| \geq t_{1-\alpha/(2*m),n-p-2}\right) \\
&\leq \sum_{i=1}^{m} P\left(|T_i| \geq t_{1-\alpha/(2*m),n-p-2}\right) \\
&= \sum_{i=1}^{m} \frac{\alpha}{m} = \alpha.
\end{aligned}
$$

Diagnostic plots

Problems in the regression function

- True regression function may have higher-order non-linear terms, polynomial or otherwise.
- We may be missing terms involving more than one $\boldsymbol{X}_{(\cdot)}$, i.e. $\boldsymbol{X}_i \cdot \boldsymbol{X}_j$ (called an *interaction*).
- Some simple plots: *added-variable* and *component plus residual* plots can help to find nonlinear functions of *one variable*.

Diagnostic plots

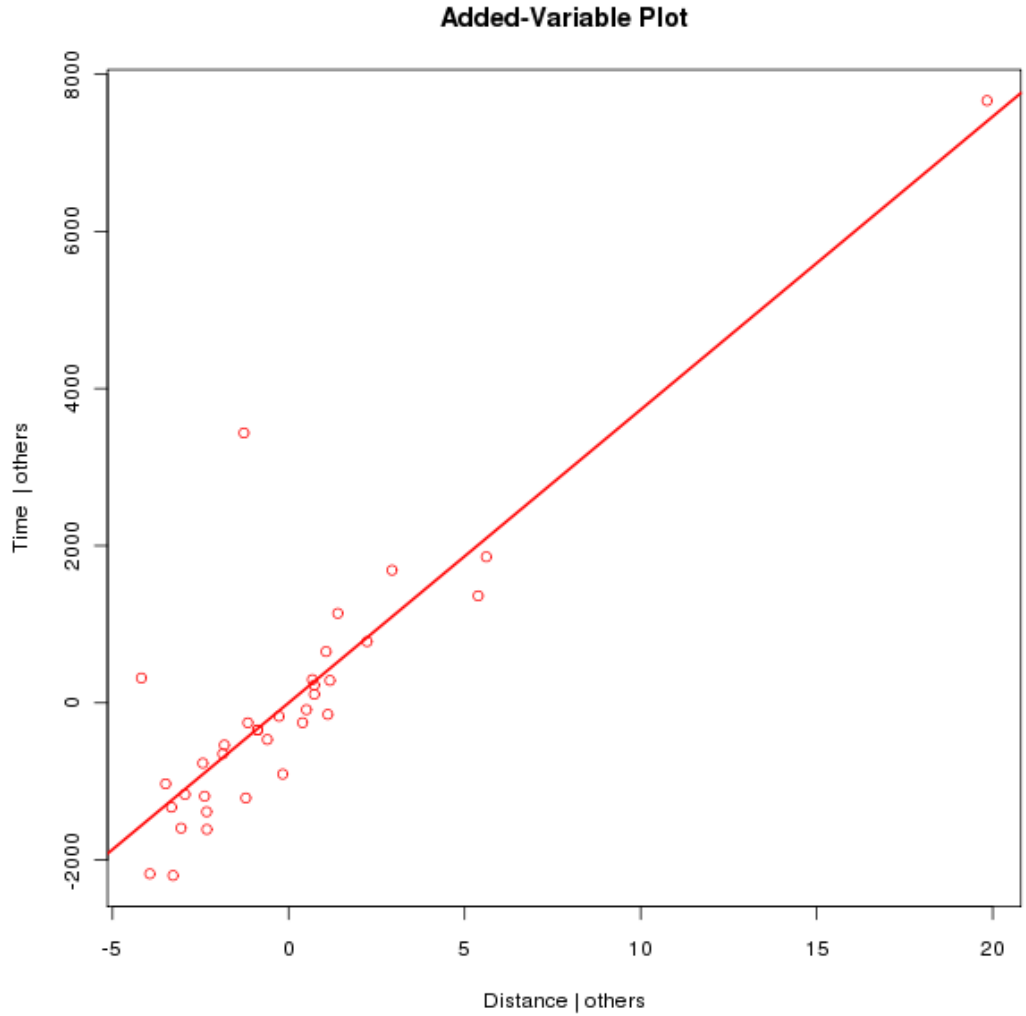
Added variable plots

- Useful for finding influential points, outliers.
- Procedure:
  - let $\tilde{e}_{X_j,i}, 1 \leq i \leq n$ be the residuals after regressing $\boldsymbol{X}_j$ onto all columns of $\boldsymbol{X}$ except $\boldsymbol{X}_j$;
  - let $e_{X_j,i}$ be the residuals after regressing $\boldsymbol{Y}$ onto all columns of $\boldsymbol{X}$ except $\boldsymbol{X}_j$;
  - Plot $\tilde{e}_{X_j}$ against e_{X_j}.

Added variable: `Distance`

Statistics 191: Introduction to Applied Statistics

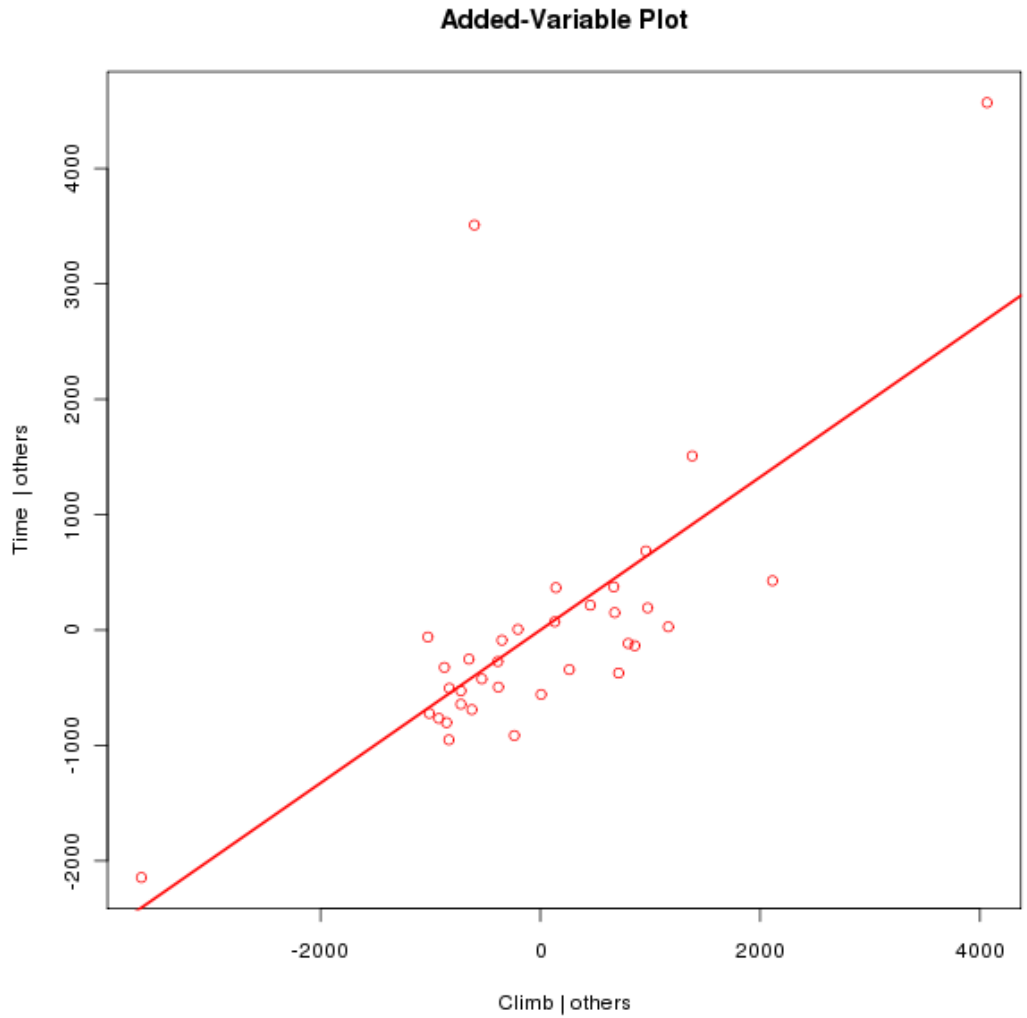
Jonathan Taylor
Department of Statistics
Stanford University



Added variable: `Climb`

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Diagnostic plots

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University

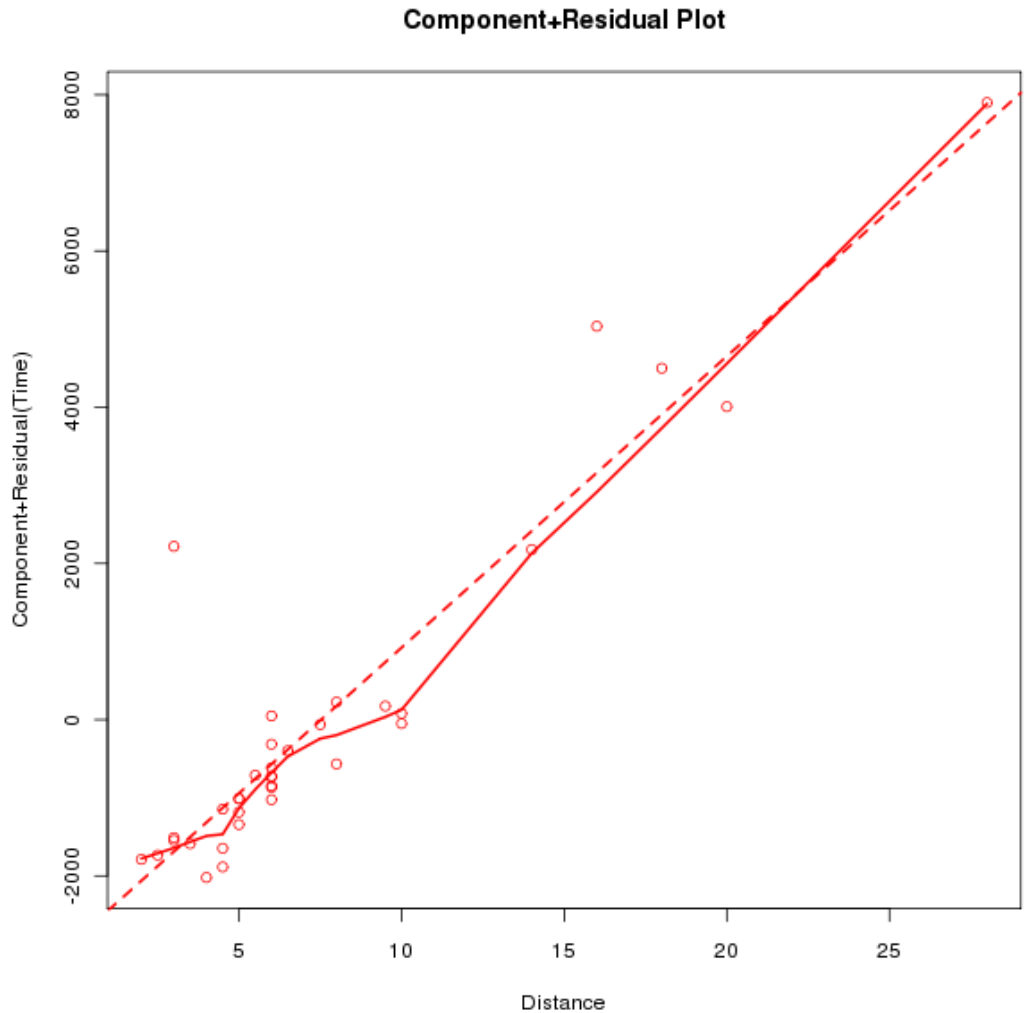
Component + residual plots

- Can help to determine non-linear trend in data.
- Procedure: plot $X_{ij}, 1 \leq i \leq n$ vs. $e_i + \widehat{\beta}_j \cdot X_{ij}, 1 \leq i \leq n$.

Component + residual: `Distance`

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Component + residual: `Climb`

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University