SMIII lectures

Week 4

The marginality principle

The marginality principle

Whenever an interaction term is included in the model, all implied lower order interactions and main effects must also be included.

Example

Suppose the interaction term `A:B:C` is to be included in a model.

The marginality principle requires that all of the following model terms must also be included:

- the two-way interaction terms, `A:B`, `A:C`, `B:C`,
- the main effects `A`, `B`, `C`,
- and the intercept `1`.

Why do we need the marginality principle?

- It is necessary to impose constraints on the parameters of the model to achieve identifiability;
- The constraints will not affect the model provided that the marginality principle is observed.

The marginality principle in R

The standard interaction operator $*$ expands according to the marginality principle:

$$A * B = 1 + A + B + A : B$$

If terms are explicitly removed from a model formula in R, the corresponding constraints are also removed.

Example

If the `A` main effect is removed (in violation of the marginality principle) R compensates by removing the constraint $\gamma_{i1} = 0$.

Example

If the interaction term `A:B` is specified in isolation, i.e. without the necessary main effects, R tries to compensate by removing all constraints on γ_{ij}.

Note the columns of the resulting model matrix are **not** linearly independent!

Models with factors and covariates

The marginality principle also applies to models with factors and covariates such as,

$$A * x = A + x + A : x$$

The usual rules for factors are applied to any terms containing factors.

Polynomial regressions

A weaker version of the marginality principle is usually applied to polynomial regression models.

In particular, it is usually argued that we should not include the cubic term x^3 in a polynomial regression unless the lower order terms 1, x and x^2 are also included.

This requirement is not driven by considerations of identifiability.

Models such as $\eta = \beta x^3$ are unambiguously defined but tend not to be very useful in allowing for curvature.

They may, however, arise from certain physical models.

Assumption Checking

Assumption Checking

An important part of good statistical practice is assumption checking.

When a model is applied to data, it is important to understand that the use of the model involves making certain assumptions about the data.

If the assumptions are not valid, then conclusions based on the model have the potential to be misleading.

It is therefore important to check the assumptions as far as possible before using a statistical model to make conclusions.

Regression Assumptions

The linear regression model can be written succinctly as

$$\boldsymbol{y} = X\beta + \boldsymbol{e}$$

where $e_1, e_2, \ldots, e_n$ are IID $N(0, \sigma^2)$ **realisations**.

The assumptions can be expanded as follows:

1. The e_i all have mean zero;
2. The e_i all have variance σ^2;
3. The e_i are normally distributed;
4. The e_i are independent.

In addition, it is assumed:

5. The x variables are known without error;
6. The errors e_i are not correlated with the x variables.

Regression assumptions

In practice, most attention is focused on using the data to investigate the plausibility of assumptions 1-3.

Assumptions 4-6, are more fundamental but cannot be verified from the observed data.

It is important to understand the context in which the data arose in order to determine whether these assumptions are reasonable.

Residuals

Model checking in linear models is usually based on analysis of the **residuals**. The **ordinary residuals** are defined by

$$\hat{e}_i = y_i - \hat{\eta}_i$$

or, in vector notation,

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{\eta}} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}.$$

Residuals

The intuitive justification for consideration of residuals is that if the model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e},$$

where $e_1, e_2, \ldots, e_n$ are independent $N(0, \sigma^2)$ realisations holds, then we should have

$$\hat{e}_i \approx e_i.$$

In this case $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$ should look roughly like a sample of independent $N(0, \sigma^2)$ observations.

Distribution of residuals

A more careful analysis can be applied to the distribution of the residuals.

Suppose the model

$$\boldsymbol{Y} = X\beta + \boldsymbol{E} \text{ with } \boldsymbol{E} \sim N_n(\boldsymbol{0}, \sigma^2 I)$$

holds and consider

$$\hat{\boldsymbol{E}} = \boldsymbol{Y} - X\hat{\beta} = (I - H)\boldsymbol{Y}$$

where

$$H = X(X^T X)^{-1} X^T.$$

Distribution of residuals

It follows that $E(\hat{\boldsymbol{E}}) = \mathbf{0}$ and

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{E}}) &= \mathrm{Var}((I - H)\boldsymbol{Y}) \\ &= (I - H)\,\mathrm{Var}(\boldsymbol{Y})(I - H)^T \\ &= \sigma^2(I - H).\end{aligned}$$

This shows that unlike an IID $N(0, \sigma^2)$ sample, the residuals are typically not independent and do not have the same variances.

In particular,

$$\mathrm{Var}(\hat{E}_i) = \sigma^2(1 - h_{ii})$$

where h_{ii} is the ith diagonal element of H.

Example

```
X <- matrix(
  c(
    1, 1,
    1, 3,
    1, 5,
    1, 7
  ),
  byrow = TRUE, ncol = 2
)
X
```

```
##      [,1] [,2]
## [1,]    1    1
## [2,]    1    3
## [3,]    1    5
## [4,]    1    7
```

Example

```
H <- X %*% solve(t(X) %*% X) %*% t(X)
H
```

```
##      [,1] [,2] [,3] [,4]
## [1,]  0.7  0.4  0.1 -0.2
## [2,]  0.4  0.3  0.2  0.1
## [3,]  0.1  0.2  0.3  0.4
## [4,] -0.2  0.1  0.4  0.7
```

Example

```
1 - diag(H)
```

```
## [1] 0.3 0.7 0.7 0.3
```

Standardized Residuals

Definition 1.5

To correct for the unequal variances, the **standardized residuals** are defined by

$$\hat{e}_i' = \frac{\hat{e}_i}{s_e\sqrt{1-h_{ii}}}.$$

□

The standardized residuals thus have equal variance but are still not independent.

Studentized Residuals

A second difficulty with the raw and standardized residuals:

- an aberrant y_i-value may influence the corresponding fitted value, $\hat{\eta}_i$, and
- this may lead to an unremarkable residual,
- thus masking the fact that the data point was in fact aberrant.

Studentized Residuals

Definition 1.6

To avoid this difficulty, the **studentized residuals** are defined by

$$\hat{e}_i^* = \frac{y_i - \hat{\eta}^{(i)}}{\sqrt{\hat{\mathrm{Var}}(Y_i - \hat{\eta}^{(i)})}}$$

where

$\hat{\eta}^{(i)}$ and $\hat{\mathrm{Var}}(Y_i - \hat{\eta}^{(i)})$ are calculated from the data with the ith observation omitted.

□

Studentized Residuals

It can be shown that

$$\hat{e}_i^* = \hat{e}_i' \left(\frac{n-p-\hat{e}_i'^2}{n-p-1} \right)^{-1/2}.$$

Model checking with residuals

It is recommended that the studentized residuals be used for basic model checking as introduced in previous courses.

- The $\hat{e}_i^*$ should be plotted against the fitted values to check for lack of fit and heteroscedasticity;
- The $\hat{e}_i^*$ should be plotted against each of the predictor variables separately to check for lack of fit and heteroscedasticity;
- If an additional variable such as "time" is recorded but not considered for inclusion in the model, it is good practice to plot the residuals against time to check for any trends etc.
- If there are no problems with lack of fit, heteroscedasticity and trends over time (when applicable), a normal quantile plot of the residuals can be used to assess the assumption of normality.

Influence diagnostics

Influence diagnostics

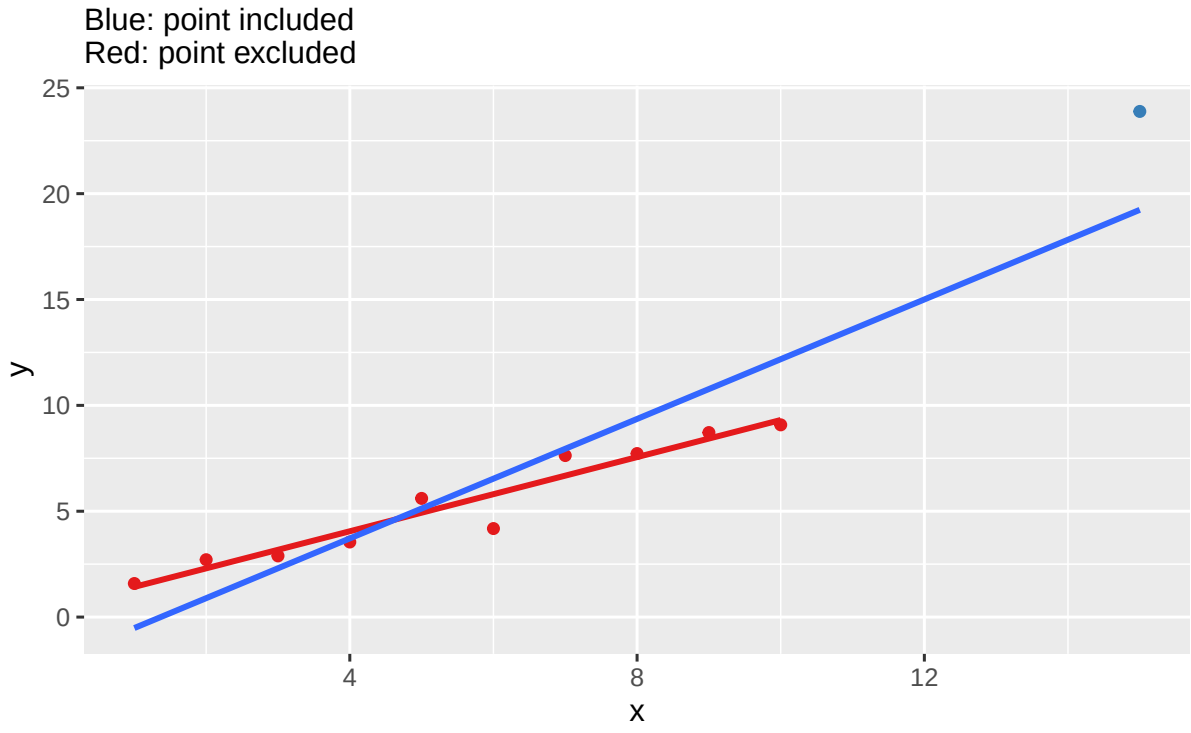
- It sometimes happens that a data set contains a small number of outliers or points with large residuals.
- The occurrence of such points can be problematic in the sense that a single outlier may have an appreciable impact on the parameter estimates.
- If it cannot be verified that the data point is erroneous then there is no good basis for its omission although one may nevertheless be suspicious.
- Influence diagnostics are used to help identify points that have a disproportionately large impact on the estimates.

Influence

- In general, a point with a small studentized residual is unlikely to have a large impact on the parameter estimates.
- A point with a large studentized residual, may or may not depending on its **leverage**.

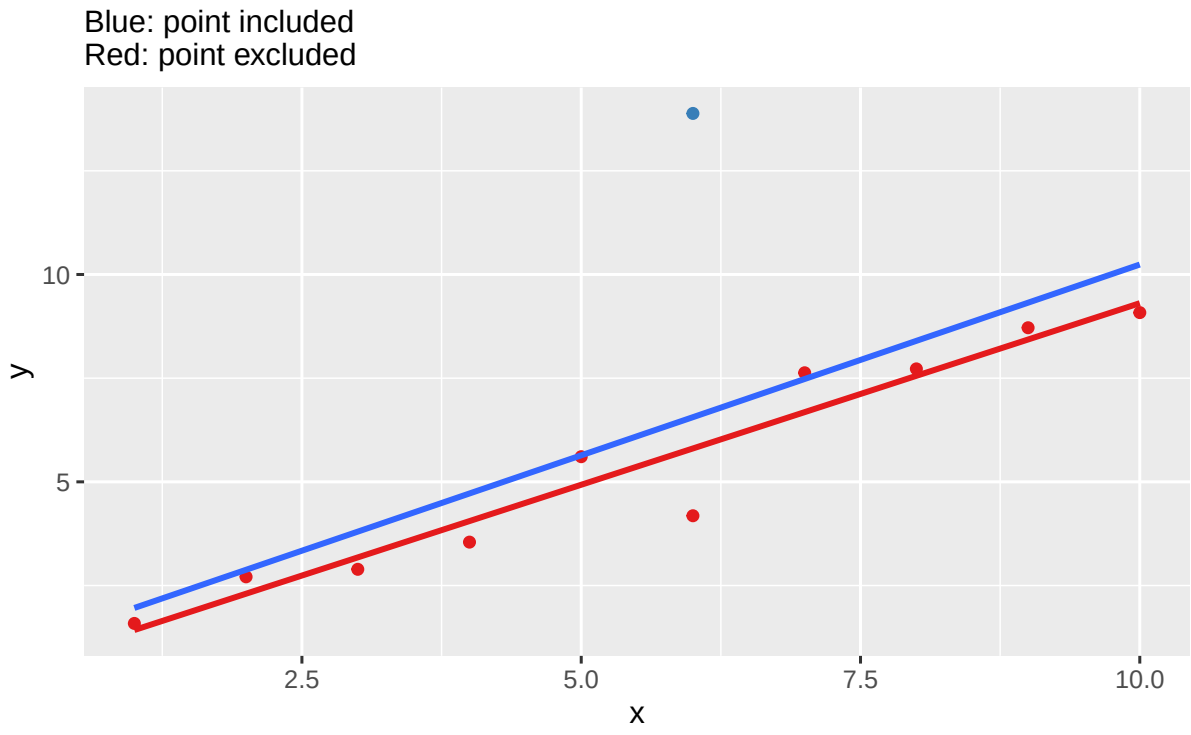
Large residual, high leverage, high influence

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



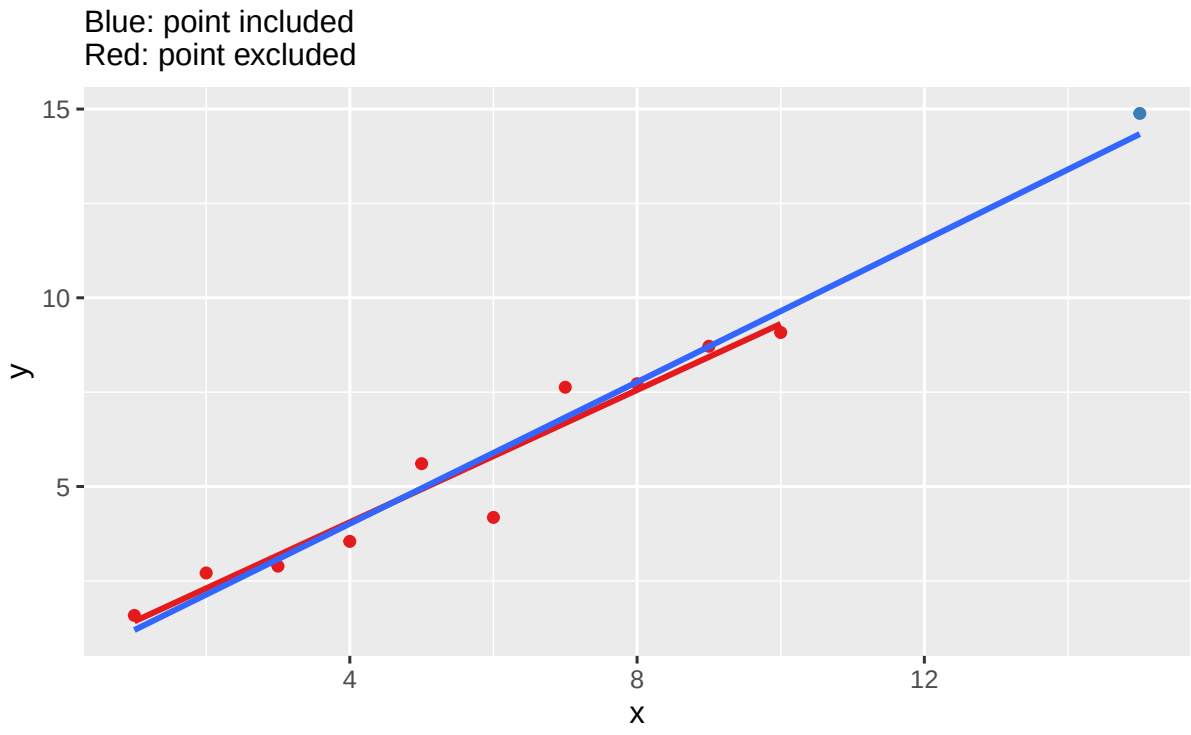
Large residual, low leverage, low influence

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



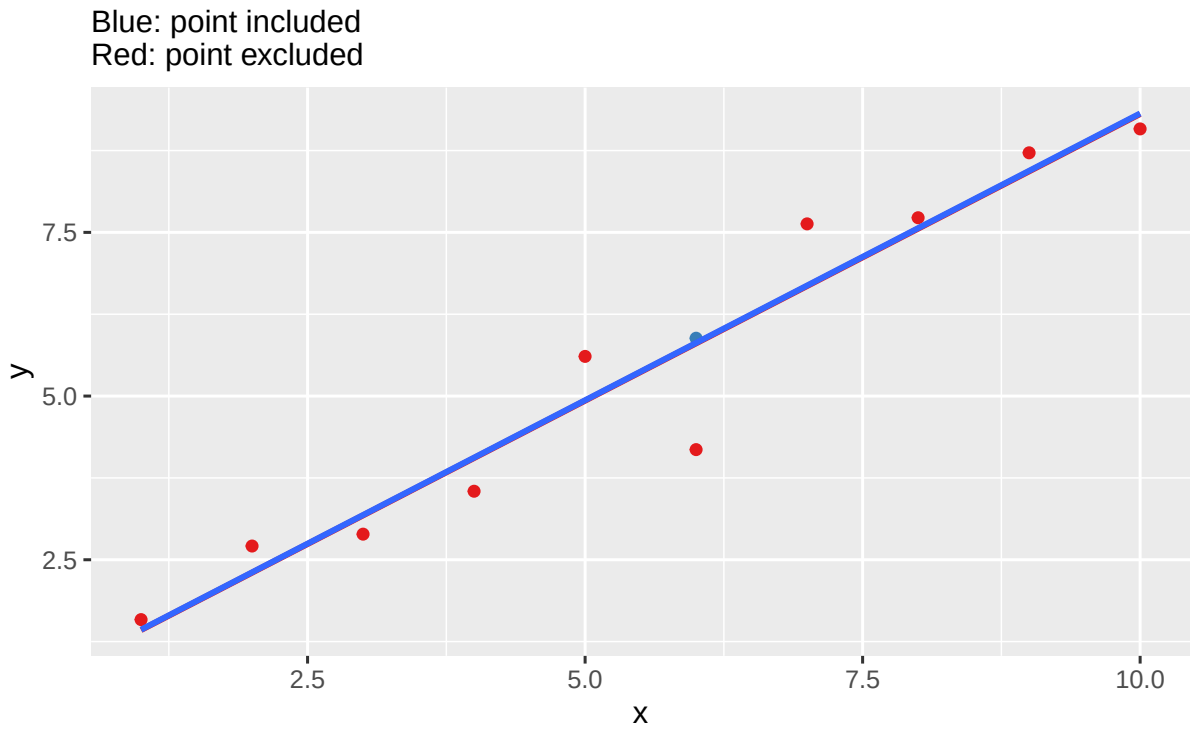
Small residual, high leverage, low influence

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



Small residual, low leverage, low influence

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



Influence and leverage

In each case, the blue line represents the least squares regression line when all points are included and the red line is the least squares regression line with the marked point excluded.

In simple linear regression, the notion of **leverage** can be thought of as a metric for the distance from x_i to $\bar{x}$.

In the first plot, where the marked point has a large residual and of high leverage, the slope is changed appreciably. In the remaining plots, the omission of the marked point has only minimal impact on the least squares regression line.

Leverage

Definition 1.7

The **leverage** of a data point is defined by

$$h_{ii} = [X(X^T X)^{-1} X^T]_{ii} = \boldsymbol{x}_i^T (X^T X)^{-1} \boldsymbol{x}_i$$

where $\boldsymbol{x}_i^T$ is the ith row of X.

□

Leverage for linear regression

It can be checked in the case of simple linear regression that

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

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Leverage

Having thus introduced the leverage, we need to define what is meant by "large". Consider

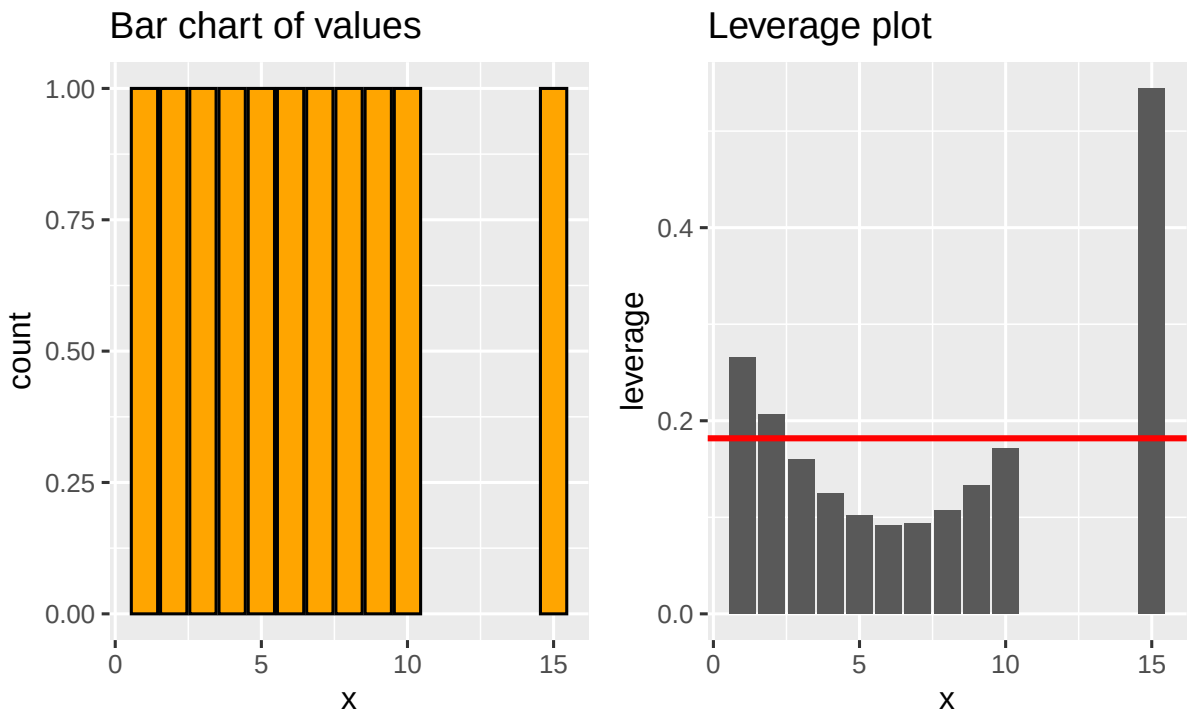
$$\begin{aligned}\sum_{i=1}^{n} h_{ii} &= \mathrm{tr}(H) \\ &= \mathrm{tr}(X(X^T X)^{-1} X^T) \\ &= \mathrm{tr}((X^T X)^{-1} X^T X) \\ &= \mathrm{tr}(I_{p\times p}) = p.\end{aligned}$$

Therefore, the average value of h_{ii} is p/n.

Leverage

- Leverages are sometimes displayed using an **index plot**.
- That is, a plot of h_{ii} vs i, to identify points of high leverage.
- Some authors suggest that points with $h_{ii} \geq 2p/n$ be highlighted.
- Shown on the following slide (for the data from the *large residual/high leverage* scatter plot) is
  - the histogram of x-values, and
  - a leverage plot.

Leverage (example)



Cook's Distance

- Points of high leverage have potentially a large impact on the estimates $\hat{\beta}$.
- Cook's distance measures how much influence each data point actually has by calculating the effect of its removal upon $\hat{\beta}$.

Cook's Distance

- Let $\hat{\beta}$ denote the least squares estimate of β based on the full data set, and
- let $\hat{\beta}^{(i)}$ be the least squares estimate when the ith data point is omitted.
- An obvious approach would be to use

$$\|\hat{\beta} - \hat{\beta}^{(i)}\|^2$$

  as a measure of the influence of the ith data point.
- However, such a measure is not satisfactory because it gives equal weight to all components of β even though their variances are generally different.
- Cook's distance uses a more appropriate distance measure.

Cook's Distance (continued)

Definition 1.8

The **Cook's distance statistic** for the ith data point is defined by

$$D_i^2 = \frac{\left(\hat{\beta} - \hat{\beta}^{(i)}\right)^T \left[\hat{\mathrm{Var}}(\hat{\beta})\right]^{-1} \left(\hat{\beta} - \hat{\beta}^{(i)}\right)}{p}$$
$$= \frac{\left(\hat{\beta} - \hat{\beta}^{(i)}\right)^T (X^T X) \left(\hat{\beta} - \hat{\beta}^{(i)}\right)}{p s_e^2}.$$

□

Cook's Distance (continued)

It can be shown that

$$D_i^2 = \frac{(\hat{e}_i')^2 h_{ii}}{p(1-h_{ii})}.$$

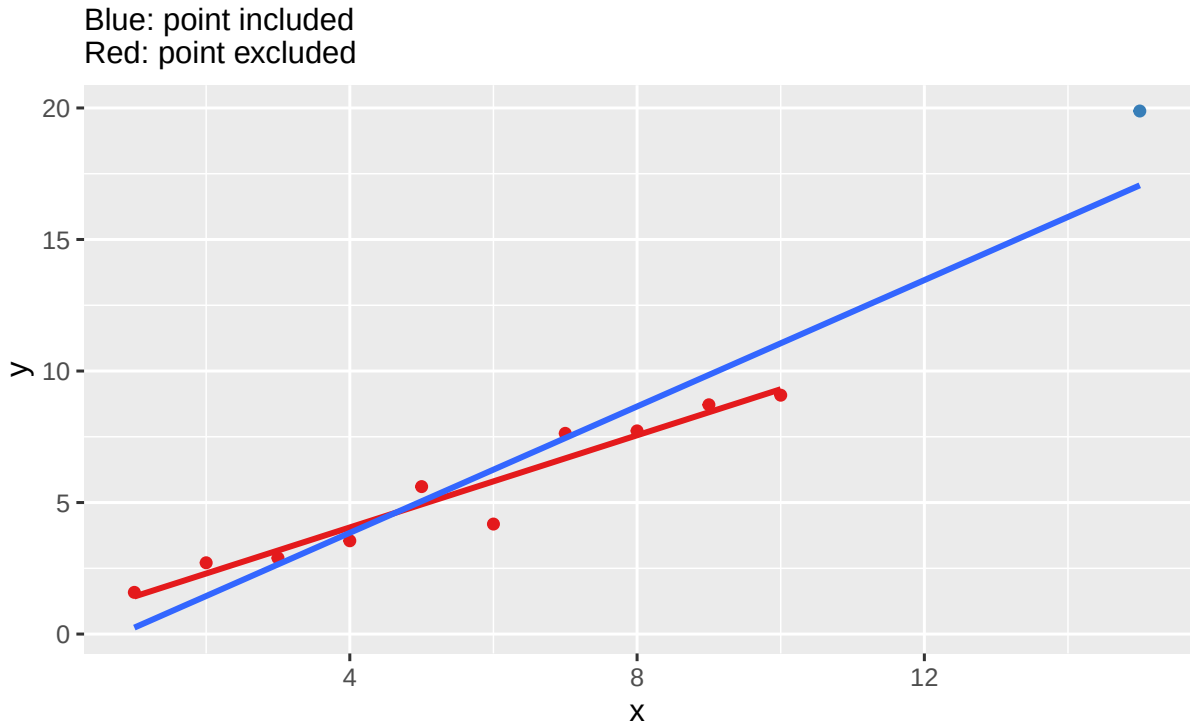
This formula for D_i^2 gives rise to the interpretation of the leverage h_{ii}.

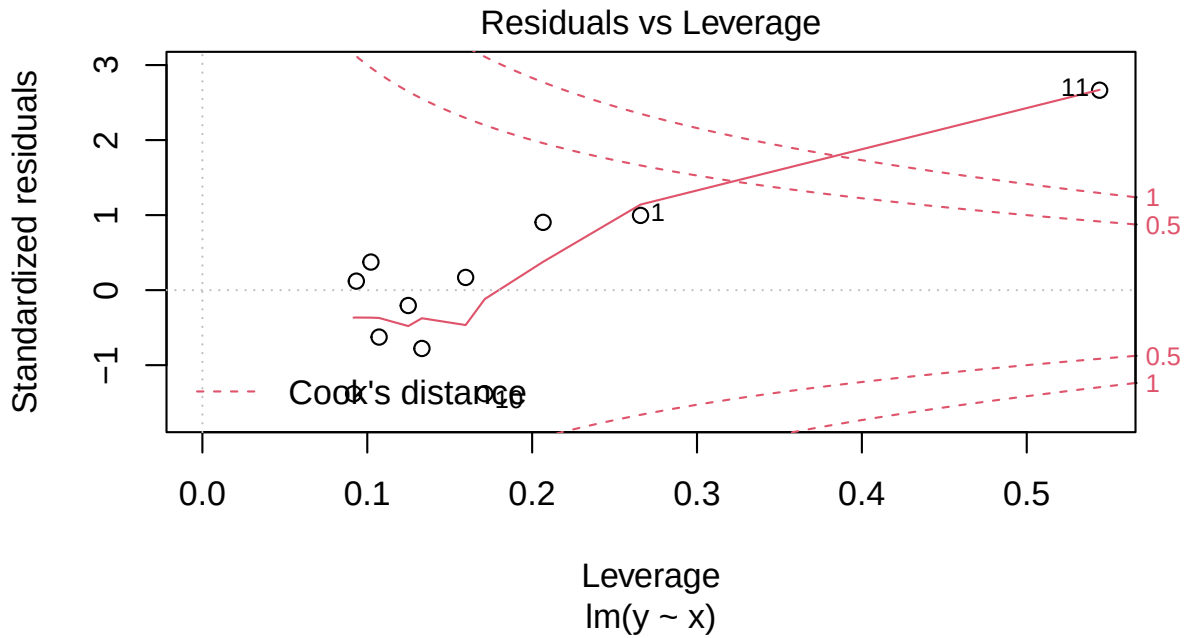
- In particular, it follows that the total impact of the ith data point on the least squares estimate, $\hat{\beta}$ will be large only if $\hat{e}_i'$ and $h_{ii}/(1-h_{ii})$ are large.
- Since $h/(1-h)$ is increasing for $0 < h < 1$, a point can be influential only when it has a high leverage h_{ii}.

Plotting Cook's Distance

- Some authors suggest that data points with $D_i^2 \geq 1$ are a cause for concern as this represents an average change to $\hat{\beta}$ of 1 standard error.
- Cook's distance can be displayed using an index plot in order to identify influential points.
- An alternative, produced in R is to plot the standardized residual $\hat{e}_i'$ vs the leverage h_{ii} and indicate the Cook's distances by contour lines.
- To illustrate, a scatter plot and the corresponding residual vs leverage plot are shown below.

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```




Residuals vs Leverage
Standardized residuals
3
2
1
0
−1
11
1
10
4
1
0.5
0.5
1
Cook's distance
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
lm(y ~ x)

Cheese example

See Case Study 3.