2.2 Linear Models (LM): estimations, residuals and assessing model assumptions

0. Introduction to Linear Gaussian Models

Linear models of the form

$$\begin{aligned} \mathbb{E}(\mathrm{Y}_i) = \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta} \\ \mathrm{Y}_i \sim N(\mu_i, \sigma^2), \end{aligned}$$

where the r.v. Y_i are independent, are the basis of most analyses of continuous data.

There are three main models of this form

- **Multivariate regression:** association between a continuous response and several explanatory variables
- **Analysis of variance (ANOVA):** comparisons of more than two means
- **Analysis of covariance (ANCOVA)**

These models (usually called **general linear model**) are usually written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.2.1}$$

where

$$\begin{aligned} \mathbf{y}^\top &= [\mathrm{Y}_1, \ldots, \mathrm{Y}_N] \\ \boldsymbol{\beta}^\top &= [\beta_1, \beta_2, \ldots, \beta_p] \\ \boldsymbol{\varepsilon}^\top &= [\varepsilon_1, \ldots, \varepsilon_N] \end{aligned}$$

where $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ for $i = 1, \ldots, N$.

The **error** term includes all the terms we have missed with the model: the true relationship is not linear, there are variables not considered, there is measurement error. The error term is usually considered **independent** from the $\mathbf{X}$.

Additionally, $\mathbf{X}$ is an $N \times p$ **design matrix**, which in the case of a multiple regression above is set to

$$\mathbf{X} = \begin{pmatrix} 1 & \mathrm{X}_{12} & \mathrm{X}_{13} & \ldots & \mathrm{X}_{1p} \\ . & & & & \\ . & & & & \\ . & & & & \\ 1 & \mathrm{X}_{N2} & \mathrm{X}_{N3} & \ldots & \mathrm{X}_{Np} \end{pmatrix}.$$

The parameter β_j is interpreted as the **average** effect on Y of a one unit increase in the covariate x_j, **holding all the other predictors fixed.**

The model **is linear in the parameters**, which means we have, for instance:

$$\mathbb{E}(\mathrm{Y}_i) = \beta_1 + \beta_2 \mathrm{X}_{i2} + \beta_3 \mathrm{X}_{i3}^2$$

$$\mathbb{E}(\mathrm{Y}_i) = \beta_1 + \gamma_1 \delta_1 \mathrm{X}_{i2} + \exp(\beta_2) \mathrm{X}_{i3}.$$

But **NOT**:

$$\mathbb{E}(\mathrm{Y}_i) = \beta_1 + \beta_2 \mathrm{X}_{i2}^{\beta_2}$$

$$\mathbb{E}(\mathrm{Y}_i) = \beta_1 \exp(\beta_2 \mathrm{X}_{i2}).$$

1. Estimation and accuracy of coefficient estimates in Linear Gaussian Models

There is exists several methods to estimate the coefficients of a linear (Gaussian) model.

In this section we provide details on how to perform

1. **Maximum likelihood estimation:** we use the distributional assumptions to derive the likelihood.
2. **Least squares estimation:** we don't make any further assumptions about the distribution of the response variable Y.

We will quantify the uncertainty that comes with the estimation by constructing confidence intervals.

1.1 Maximum likelihood estimation

The score function is given by

$$U_j = \sum_{i=1}^{N} \left[\frac{(y_i - \mu_i)}{\mathbb{V}ar(Y_i)} \mathrm{X}_{ij} \left(\frac{d\mu_i}{d\eta_i} \right) \right]$$

while the information is of the form

$$\mathcal{I} = \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{X}$$

If we want to apply the **method of scoring** to approximate the MLE, the estimating equation is

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(m)} &= \hat{\boldsymbol{\beta}}^{(m-1)} + \left[\mathcal{I}^{(m-1)} \right]^{-1} \mathbf{u}^{(m-1)} \\ \left[\mathcal{I}^{(m-1)} \right] \hat{\boldsymbol{\beta}}^{(m)} &= \left[\mathcal{I}^{(m-1)} \right] \hat{\boldsymbol{\beta}}^{(m-1)} + \mathbf{u}^{(m-1)} \end{aligned} \tag{2.2.4}$$

From Equation $(2.2.2)$ (in the proof above), the information matrix can be written as

$$\mathcal{I} = \mathbf{X}^\top \mathbf{W} \mathbf{X} \tag{2.2.5}$$

where $w_{ii} = \frac{1}{\mathbb{V}\mathrm{ar}(\mathrm{Y}_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2 = \frac{1}{\sigma^2}$

Finally, the expression on the right hand side of $(2.2.4)$ can be written as

$$\sum_{k=1}^{p} \sum_{i=1}^{N} \frac{\mathrm{X}_{ij}\mathrm{X}_{ik}}{\mathbb{V}\mathrm{ar}(\mathrm{Y}_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2 \hat{\beta}_k^{(m-1)} + \sum_{i=1}^{N} \frac{(\mathrm{Y}_i - \mu_i)\mathrm{X}_{ij}}{\mathbb{V}\mathrm{ar}(\mathrm{Y}_i)} \left(\frac{d\mu_i}{d\eta_i} \right)$$

which can be written in matrix terms as

$$\mathcal{I}^{(m-1)} \hat{\boldsymbol{\beta}}^{(m-1)} + \mathbf{u}^{(m-1)} = \mathbf{X}^\top \mathbf{W} \mathbf{z} \tag{2.2.6}$$

where

$$\mathrm{Z}_i = \sum_{k=1}^{p} \mathrm{X}_{ik}\hat{\beta}_k^{(m-1)} + (\mathrm{Y}_i - \mu_i)\left(\frac{d\eta_i}{d\mu_i}\right) = \sum_{k=1}^{p} \mathrm{X}_{ik}\hat{\beta}_k^{(m-1)} + \left(\mathrm{Y}_i - \sum_{k=1}^{p} \mathrm{X}_{ik}\beta_k^{(m-1)}\right) = \mathrm{Y}_i$$

And, therefore,

$$\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}\hat{\beta} = \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{y} \quad\Longrightarrow\quad \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Properties:

- Unbiasedness

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}(\mathrm{Y}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

- The **variance-covariance** matrix is $\mathcal{I}^{-1}$, therefore

$$\mathcal{I}^{-1} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

- Normality

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$$

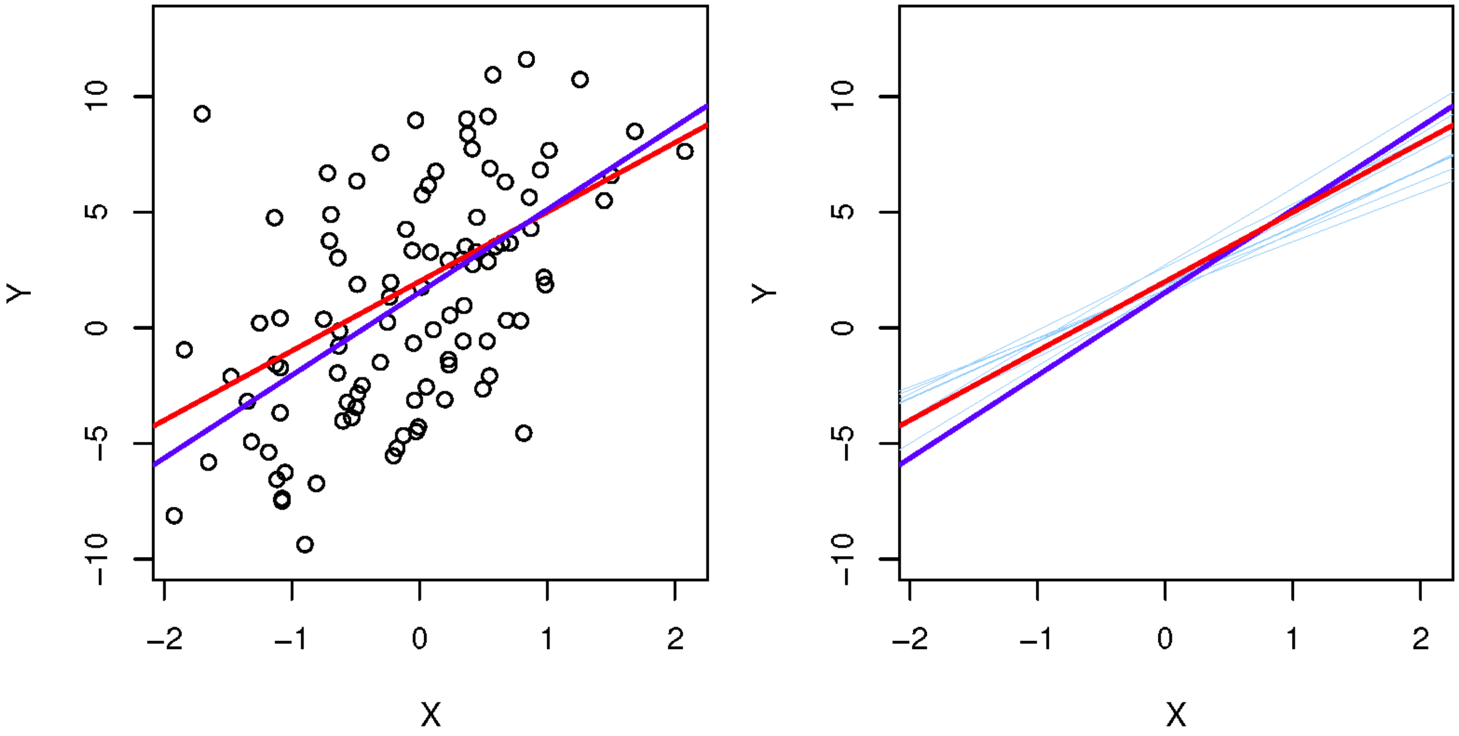


Figure 2.2.1: Left: Simulated data. Red line gives the true relationship (population regression line), blue line gives the least square estimate. Right: light blue lines are 10 least square lines based on

random samples.

1.2 Least squares estimation

It is possible to derive an estimator without making any further assumption about the distribution of $\mathbf{y}$.

Under **Gauss-Markov assumptions**, i.e.

- $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$
- $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \mathbb{I}_N$

and assuming that $N > p$, the **least squares function is**

$$\mathrm{RSS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} \tag{2.2.7}$$

This is a multivariate function in $\boldsymbol{\beta}$ which is (strictly) **convex**.

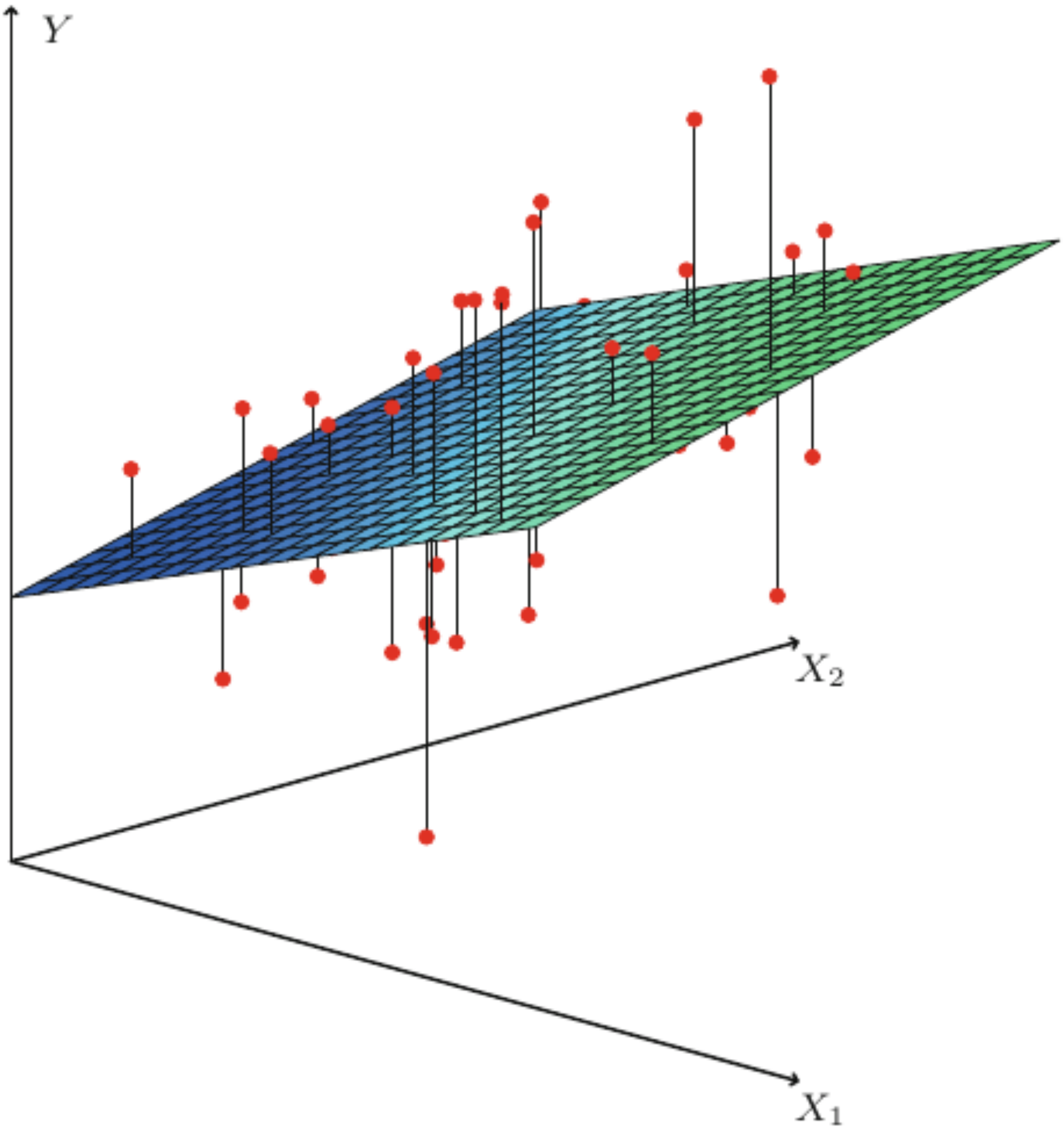
Hence there is a unique minimiser $\hat{\boldsymbol{\beta}}$, satisfying

$$\begin{aligned}\frac{d}{d\boldsymbol{\beta}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) &= \frac{d}{d\boldsymbol{\beta}}\mathbf{y}^\top\mathbf{y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \\ &= -2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = 0 \\ &\rightarrow \mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{y}\end{aligned}$$

Assuming that $\mathbf{X}$ has rank $N \geq p$, we can write

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

While this estimator is unbiased, it is necessary to introduce an assumption on the distribution of the $\mathbf{y}$ to derive the distribution of the estimator.



Y
X_2
X_1

1.3 Confidence Intervals for regression parameters

The estimate of the uncertainty of the estimates is given by the **standard error**, $\mathrm{SE}(\hat{\boldsymbol{\beta}})$:

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{\mathrm{X}}^2}{\sum_{i=1}^{N} (\mathrm{X}_i - \bar{\mathrm{X}})^2} \right]$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{N} (\mathrm{X}_i - \bar{\mathrm{X}})^2},$$

where $\sigma^2 = \mathrm{Var}(\varepsilon)$.

Similarly in the presence of other covariates.

Remark: the $\mathrm{SE}(\hat{\beta}_1)$ is smaller when the X_i are more spread out: intuitively, we are more able to estimate the slope of the line in this case.

Standard errors can be used to compute $(1-\alpha)100\%$ **confidence intervals** as:

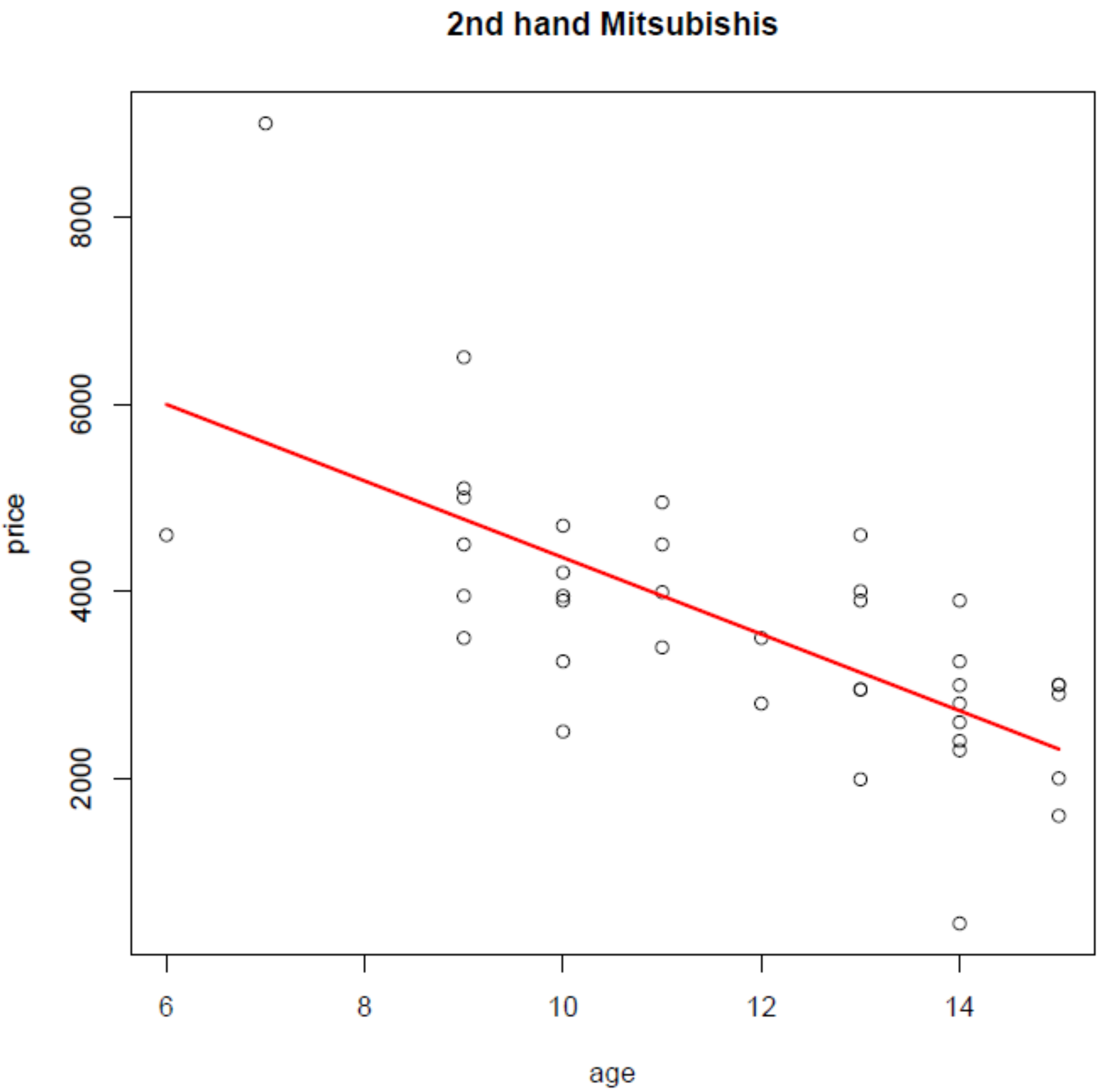
$$\left[\hat{\beta}_k - t_{1-\alpha/2,n-2} \mathrm{SE}\left(\hat{\beta}_k\right), \hat{\beta}_k + t_{1-\alpha/2,n-2} \mathrm{SE}\left(\hat{\beta}_k\right) \right],$$

$k = 0, 1$, where $t_{1-\alpha/2,n-2}$ represents the $1-\alpha/2$ quantile of the Student-t distribution with $n-2$ degrees of freedom.

1.4 Example

Consider the simple linear model $\mu_i \sim \beta_0 + \beta_1 \mathbf{age}$ where $Y_i \sim N(\mu_i, \sigma^2)$ for the price of 2nd-hand Mitsubishi car dataset. The $glm()$ (or $lm()$) function in R can be used to fit the simple regression model using the **age** as the predictor and the **price** as the response.

```r
mitsub <- read.table('/course/data/mitsub.txt', header=T)
t(head(mitsub))
attach(mitsub)
mitsub.fit <- glm(formula = price ~ age, family = gaussian())
plot(age,price,main="2nd hand Mitsubishis")
lines(age,mitsub.fit$fitted.values, col=2, lwd=2)
```



Recall what assumptions have been made about the data:

- the population mean follows a straight line:

$$\text{price} = \beta_0 + \beta_1 \text{age}$$

  for some numbers β_0 and β_1;

- in each vertical strip the prices have a normal distribution with the same (σ^2);
- the prices are independent of one another.

Data set

mitsub.txt

Activity in R: Multiple Linear Regression in R

Multiple Linear Regression can be fitted in R using the $lm()$ or the $glm()$function.

Obtain the **`basketball`** data from our **`data`** folder. The basketball data frame contains for a number of NBA basketball players the variables `PPM` (points per minute), `APM` (assists per minute), **`Height`** (height of the player), `MPG` (minutes played per game), **`Age`** (age of the player) and **`Name`** (name of the player).

- Assume `PPM` as the response and **`Height`**, `APM`, `MPG` and **`Age`** as predictors. Fit the multiple linear regression using the $lm()$ and $glm()$ functions and compare the outputs of these two functions.

- Write down the regression equation and discuss the adequacy of the coefficient estimates.

- Your friend notices that the coefficient of `MPG` is positive. He says that a player's `PPM` statistic can be increased by increasing the predictor `MPG`, that is, by giving him more play time. Do you agree? Why?

2. Distribution of residuals in Linear Gaussian Models

Let us now introduce the hat matrix

$$\boldsymbol{H} := \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top.$$

The hat matrix puts the hat on $\boldsymbol{y}$ in the following way:

Recall that $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$ and $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$

$$\boldsymbol{H}\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{y}}.$$

$\boldsymbol{H}$ is symmetric and idempotent. Therefore, it is a projection matrix. It defines a transformation of $\boldsymbol{y}$ in N-dimensional space to vector $\hat{\boldsymbol{y}}$ in a subspace such that $\hat{\boldsymbol{y}}$ is as close to $\boldsymbol{y}$ as possible.

Note also that for an indempotent matrix $\boldsymbol{A}$: $\operatorname{rank}(\boldsymbol{A}) = \operatorname{tr}(\boldsymbol{A})$.

A fitted value has the form

$$\hat{y}_i = \sum_{j=1}^{n} H_{ij} y_j,$$

a weighted sum of the y_j's . Thus the effect that y_i has on its fitted value is H_{ii} the ith diagonal entry of H. The ith diagonal of the hat matrix H_{ii} will give us the so-called leverage point, which will help us with diagnosing influential points in the regression (See later).

Recall that the fitted values are $\hat{\boldsymbol{y}} := \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Then the residuals are given by

$$\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}. \qquad (2.2.8)$$

Theorem

If the Gauss-Markov assumptions hold, then

$$r \sim \mathcal{N}\left(0, \sigma^2(\mathbb{I} - \mathrm{H})\right)$$

Theorem

Under the Gauss-Markov assumptions,

$$\sigma^{-2}\mathbf{r}^\top\mathbf{r} = \sigma^{-2}\sum_{i=1}^{N} r_i^2 \sim \chi^2_{N-p}$$

Under the Linear Gaussian Model assumptions,

$$\sigma^{-2}\boldsymbol{r}^\top\boldsymbol{r} = \sigma^{-2}\sum_{i=1}^{N} r_i^2 \sim \chi^2_{(N-p)}.$$

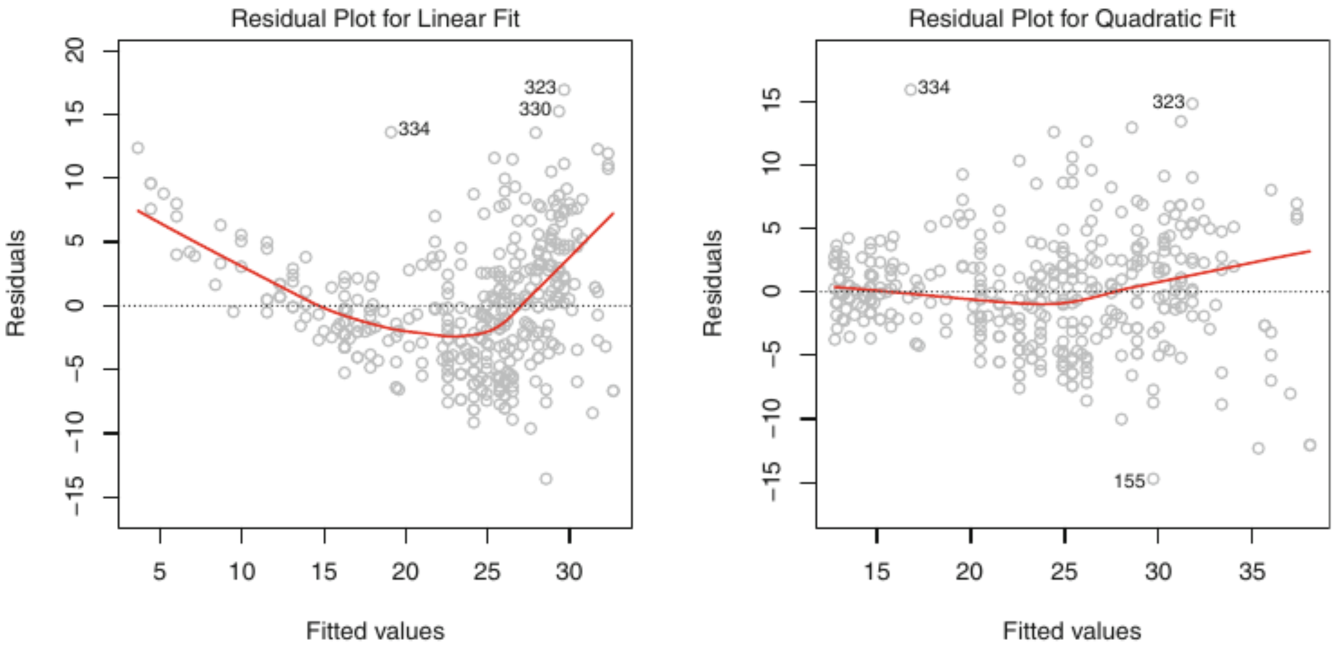
It follows that $\hat{\sigma}^2 := \sum_{i=1}^{N} r_i^2/(N-p)$ is an unbiased estimator of σ^2. (Proof omitted). Note that for the simple linear model $p = 2$ (number of parameters in the model).

This is called the **residual standard error** and it is the estimate needed to estimate the coefficient standard error.

Residual plots are useful graphical tools for identifying non-linearity in the data: we can plot the residuals $(\mathrm{Y}_i - \hat{\mathrm{Y}}_i)$ versus the fitted values $\hat{\mathrm{Y}}_i$.

Ideally the residual plot will not show any discernible pattern.

If the residual plot indicates that there are non-linear associations in the data, then a simple approach is to use non-linear transformations of the predictors, i.e. $\log x$, $\sqrt{x}$, x^2, etc.



3. Assessing model assumptions

We have made several assumptions for the model to be valid. It is therefore needed to check if these assumptions hold. In order to do so we look into:

1. The standardised residuals
2. The presence of high leverage points
3. The Cook's distance

3.1 Standardised residuals

The residuals vs fitted values plot may reveal possible violations of linearity or homoscedasticity. Standardising the residuals may lead to a better feeling for their magnitude.

We have calculated the variance of the i-th residual to be

$$\mathrm{Var}[r_i] = \mathrm{Cov}[\boldsymbol{e}_i^\top \boldsymbol{r}] = \boldsymbol{e}_i^\top \sigma^2 (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{e_i} = \sigma^2 (1 - h_{ii}). \tag{2.2.9}$$

Accordingly, one defines the standardised residuals:

$$r_{0i} = \frac{r_i}{\sqrt{\hat{\sigma}^2 (1 - h_{ii})}}.$$

Recommendations about outliers:
Points should not be routinely deleted from an analysis just because they do not fit the model. Outliers and bad leverage points are signals, flagging potential problems with the model. Outliers often point out an important feature of the problem not considered before. They may point to an alternative model in which the points are not an outlier.

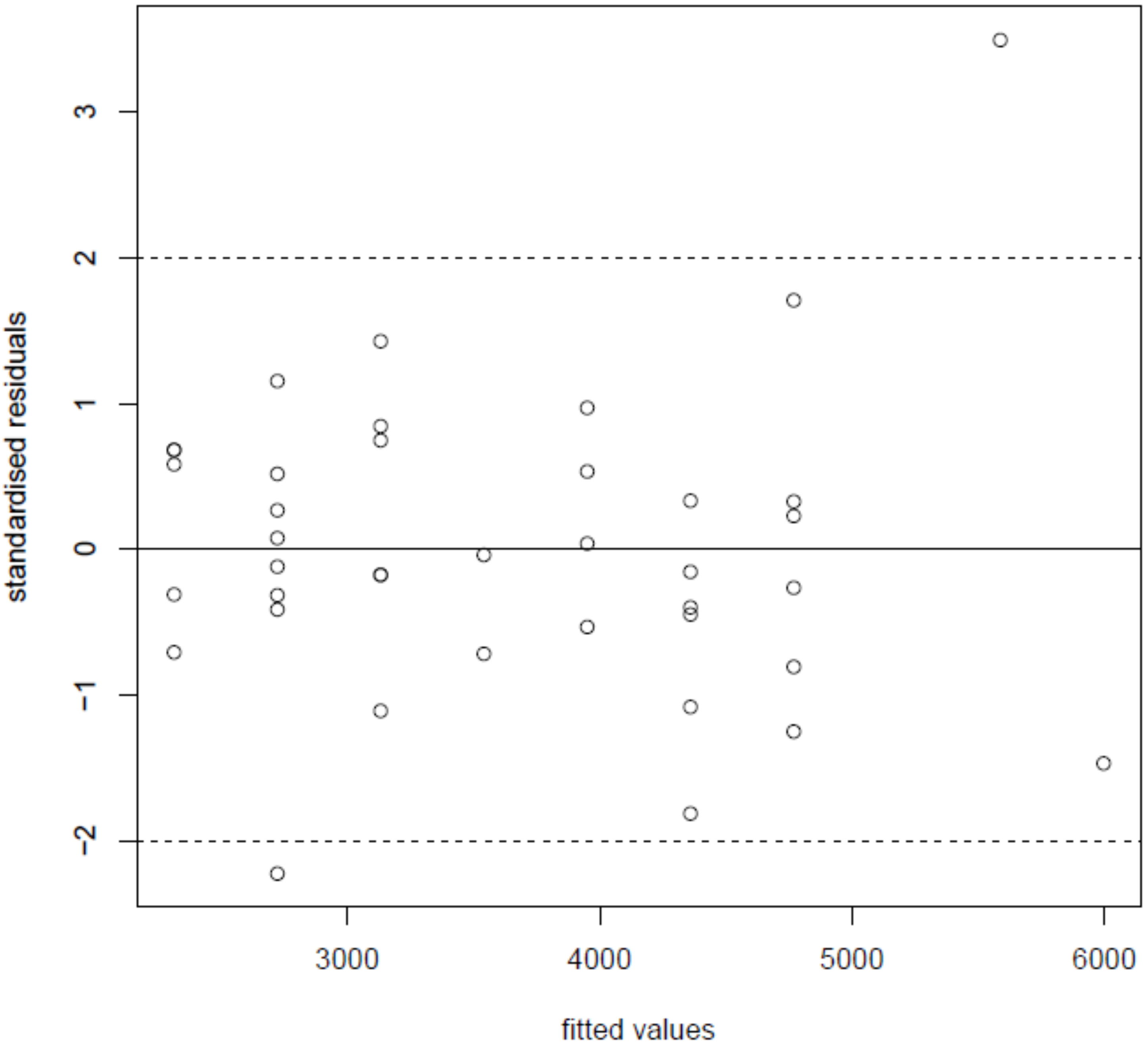
Here, the diagonal of the hat matrix may be calculated in R as below with 2 options.

```
mitsub <- read.table('/course/data/mitsub.txt', header=T)
attach(mitsub)
mitsub.fit <- glm(formula = price ~ age, family = gaussian())
X=model.matrix(mitsub.fit)

## option 1
H=X%*%solve(t(X)%*%X)%*%t(X)
diag(H)

## option 2
hat(X)
```

To plot the standardised residuals, execute the following R code:

```
mitsub <- read.table('/course/data/mitsub.txt', header=T)
attach(mitsub)
mitsub.fit <- glm(formula = price ~ age, family = gaussian())
X=model.matrix(mitsub.fit)
H=X%*%solve(t(X)%*%X)%*%t(X)

r=mitsub.fit$residuals
sigmahat=sqrt(sum(r^2)/mitsub.fit$df.residual)
r0=r/(sigmahat*sqrt(1-diag(H)))
yhat=mitsub.fit$fitted.values
plot(yhat,r0,xlab="fitted values",ylab="standardised residuals")
```

```
abline(h=c(-2,0,2), lty=c(2,1,2))
```



Except for the single high residual at about 5800, the residuals tend to have no particular pattern and appear as a randomly scattered equal-width band of points around the horizontal axis, in line with our model assumptions. Standardised residuals convey an idea about the magnitude of the outlier: For instance, about 95% of residuals are expected to lie within ± 2 standard deviations. The fact that the highest standardised residual is 3.3075342 indicates a violation of the normality assumption (an outlier).

In this case, the outlier was a 7-year-old car which was particularly well looked after and had some additional features. One may argue that it is not representative of "typical" Mitsubishi Sigma cars, and then omit it from the sample. This will yield a better fit.

i **Recommendations about outliers:** Points should not be routinely deleted from an analysis just because they do not fit the model. Outliers and bad leverage points are signals, flagging potential problems with the model. Outliers often point out an important feature of the problem not considered before. They may point to an alternative model in which the points are not an outlier.

3.2 Leverage

The i-th diagonal entry h_{ii} of $\boldsymbol{H}$ is called the $leverage$ of the i-th observation.

Let $\hat{y}_i^{-i}$ denote the value at x_i of the model which is fitted to the data where (x_i, y_i) is removed. Then

$$\frac{\hat{y}_i - \hat{y}_i^{-i}}{r_i} = \frac{h_{ii}}{1 - h_{ii}}$$

where $r_i = y_i - \hat{y}_i$ is the i-th residual.

That is, the leverage of the i-th observation is the difference in the i-th fitted value after removing (x_i, y_i), divided by the i-th residual. Model fits are particularly sensitive to data with high leverage. Also note from $(2.2.9)$ that residuals at points with high leverage have *small* variance.

It can be shown that for simple linear regression $y_i = \beta_0 + \beta_1 x_1$ the leverage is largest at the most extreme x-values.

The sum of leverages equals the number of parameters:

$$\sum_{i=1}^{N} h_{ii} = \operatorname{tr}(\boldsymbol{H}) = \operatorname{tr}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \operatorname{tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}) = \operatorname{tr}(\boldsymbol{I_p}) = p$$

Hence on average the leverage is p/N. As a rule of thumb, if h_{ii} is greater than two or three times p/N, it may be a concern.

To see which points x_i have high leverage,

```
mitsub <- read.table('/course/data/mitsub.txt', header=T)
attach(mitsub)
mitsub.fit <- glm(formula = price ~ age, family = gaussian())
X=model.matrix(mitsub.fit)

age[which(hat(X)>= 3*2/length(age))]
```

That is, cars with age 6 (the smallest age) have high leverage.

3.3 Cook's distance

The Cook's distance is defined by:

$$D_i = \frac{1}{p\hat{\sigma}^2}|\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}^{-i}|^2$$

Cook's distance hence measures the (rescaled) sum of squared differences between fitted values when the i-th datum is removed. It is a measure for the **influence** of the i-th datum on the entire model fit.

Activity [Optional] : Cook's distance

Show that

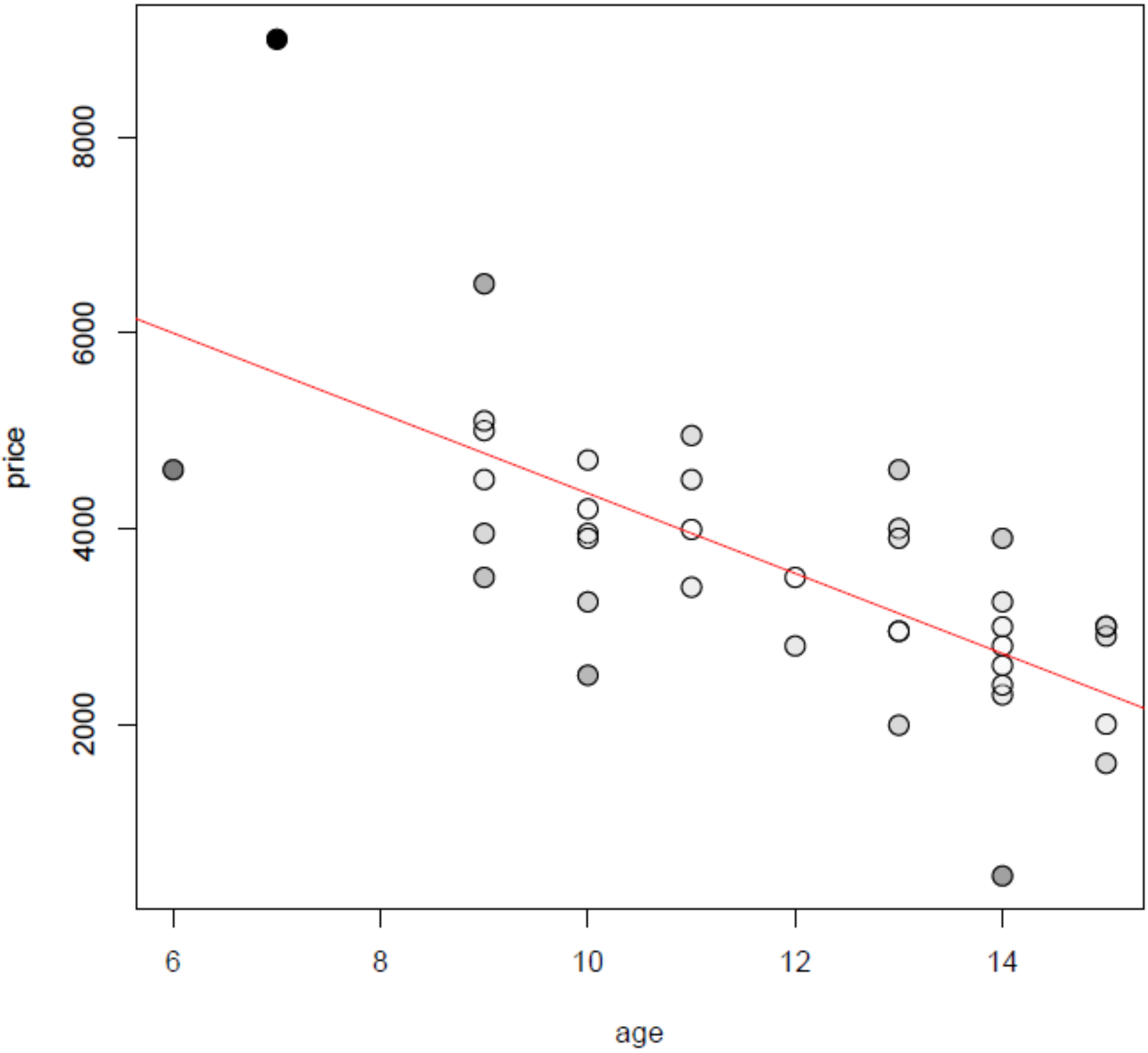
$$D_i = \frac{1}{p}\frac{h_{ii}}{1-h_{ii}}r_{0i}^2$$

Fox (2002, p. 198) is among many authors who recommend $4/(n-2)$ as a rough cutoff for noteworthy values of D_i for simple linear regression. In practice, it is important to look for gaps in the values of Cook's distance and not just whether values exceed the suggested cut-off.

Example: Cook's distance

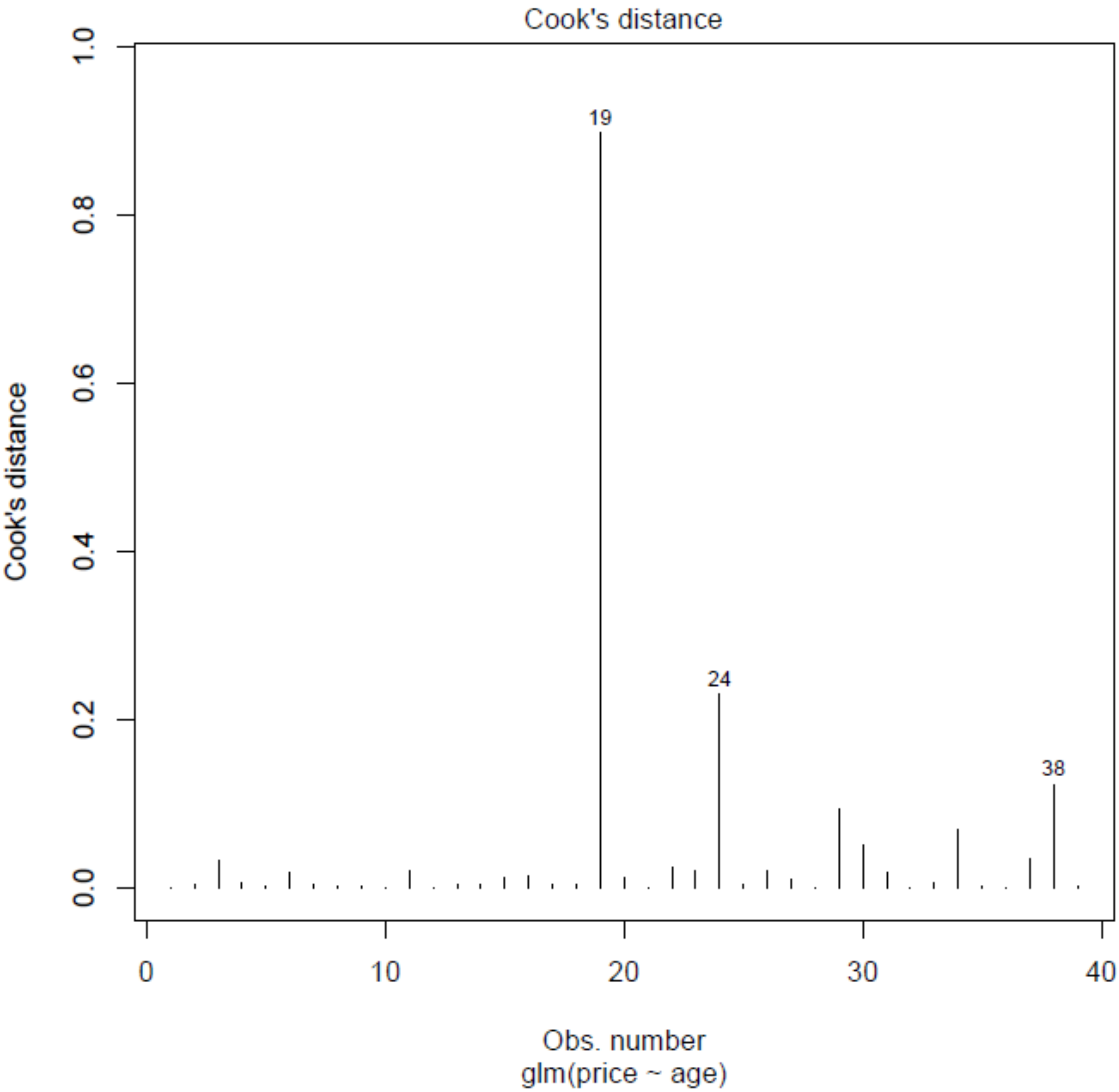
In this example, we produce a simple linear regression plot with data points shaded according to their Cook's distance.

```
mitsub <- read.table('/course/data/mitsub.txt', header=T)
attach(mitsub)
mitsub.fit <- glm(formula = price ~ age, family = gaussian())

attach(mitsub)
cookd <- cooks.distance(mitsub.fit)
cookd <- cookd/max(cookd)
cook.colours <- gray(1-sqrt(cookd))
plot(age,price,bg=cook.colours,pch=21,cex=1.5)
points(age,price,pch=1,cex=1.5)
abline(lm(price~age)$coefficients,col=2)
```



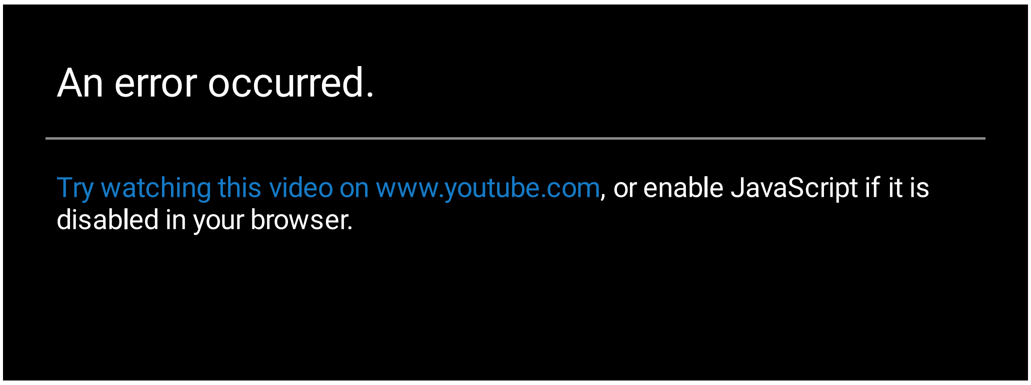
Another type of Cook's distance plot can be obtained by using the following command:

```
mitsub <- read.table('/course/data/mitsub.txt', header=T)
attach(mitsub)
mitsub.fit <- glm(formula = price ~ age, family = gaussian())

plot(mitsub.fit,which=4)
```



This plot shows that the highest Cook's distance corresponds to the 19th, 24th and 38th observation.

Leverage and Influential Points in Simple Linear Regression



An error occurred.

Try watching this video on www.youtube.com, or enable JavaScript if it is disabled in your browser.

[redacted]

Activity in R: Distribution of residuals

Activity in R: Distribution of residuals

Recall the `basketball` dataset from the previous activity. Plot the standardised residuals vs fitted values for the case of the multivariate regression from the previous activity. Interpret the obtained plot.