STAT 210
Applied Statistics and Data Analysis
Linear Regression III:
Confidence Bands and Anova

Joaquin Ortega

Fall 2020

Results from previous lectures

Results from previous lectures

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0. \tag{1}$$

Let $V = (\mathbf{X}'\mathbf{X})^{-1}$, then

$$Var(\hat{\beta}_0) = \sigma^2 v_{11}, \qquad Var(\hat{\beta}_1) = \sigma^2 v_{22} \tag{2}$$

where v_{ii} is the i-th diagonal element of V.

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n. \end{pmatrix} \tag{3}$$

Confidence Band for the Regression Line

Confidence Band for the Regression Line

So far, we have looked at confidence intervals for the parameters of the regression line, but what about confidence intervals for the values of the regression?

Recall that our model is

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$. Therefore, since X is not random,

$$E(Y) = \beta_0 + \beta_1 X$$

Thus, the value of the regression line at $X = x$ represents the average response at that point. We have denoted this value by $\hat{y}$, but to make the following argument clearer, let's change the notation to $\mu_{Y|x}$ to emphasize that we are considering the average value of the response Y at point x given by the regression model.

Confidence Band for the Regression Line

For $\mu_{Y|x}$, there are two sources of variability, $\hat{\beta}_0$, and $\hat{\beta}_1$.

The standard error (or empirical standard deviation) of $\mu_{Y|x}$ is

$$se_{\mu_{Y|x}} = \hat{\sigma}\Big(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2}\Big)^{1/2}. \tag{4}$$

Observe that the standard error is a minimum when $x = \bar{x}$.

We have shown that the regression line passes through the point $(\bar{x}, \bar{y})$, and the predicted value at $\bar{x}$ will be $\bar{y}$, whatever the slope.

When we want to make a prediction away from $\bar{x}$, we have to take into account the uncertainty in the slope of the regression line, and the confidence interval grows wider.

Confidence Band for the Regression Line

A confidence interval for the average value of Y at x at the $(1-\alpha)$ level is given by

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x - t_{n-2,1-\alpha/2} se_{\mu_{Y|x}}, \hat{\beta}_0 + \hat{\beta}_1 x + t_{n-2,1-\alpha/2} se_{\mu_{Y|x}}\right)$$

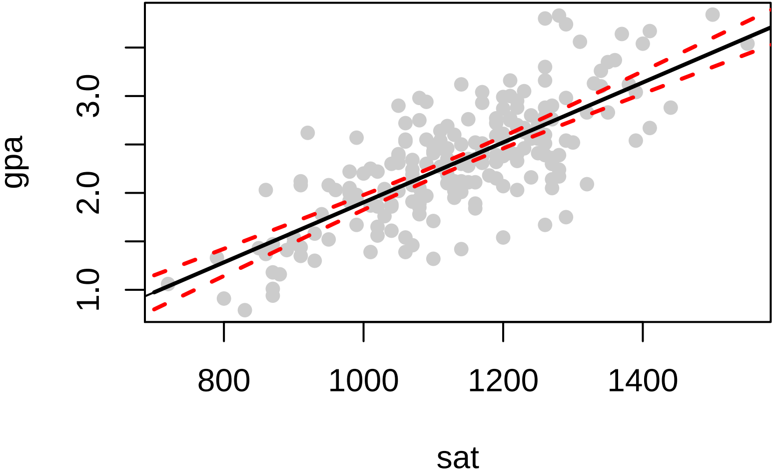
We can get these intervals using the function `predict`, which, when applied to an object of class `lm` and a data frame of x values, will give the values of the regression line at the x values, with the option of adding confidence intervals.

```
new.data <- data.frame(x=c(900,1100,1300))
predict(model,new.data,interval='c')
```

```
##        fit      lwr      upr
## 1 1.592779 1.486950 1.698609
## 2 2.211634 2.154371 2.268896
## 3 2.830488 2.746088 2.914887
```

Confidence Band for the Regression Line

Let us use this to draw 'confidence bands' for the regression line in this example.



Confidence Band for the Regression Line

```r
plot(sat, gpa)
modelA <- lm(gpa~sat, data = Grades)
abline(modelA)
new.sat <- data.frame(sat=seq(700,1600,
                              length.out = 15))
pc <- predict(modelA,new.sat, int='c')
matlines(new.sat$sat, pc, lty=c(1,2,2),
         lwd=rep(2,3),
         col=c('black','red','red'))
```

Confidence Band for the Regression Line

These confidence bands look too narrow for the uncertainty in the model but remember that they are based on confidence intervals for the (predicted) average value.

We are only taking into account the uncertainty in the estimation of the parameters of the model and not sampling variability.

If we wanted to predict the value of y corresponding to a given value of x (instead of predicting the *average* value of y at x), we would expect a wider confidence band.

Confidence Band for the Regression Line

To avoid confusion, these are called **prediction** intervals.

Prediction intervals are wider because they take into account sampling variability due to the error term in the model.

Also, since the uncertainty in the estimation of the parameters is less important, their curvature is less pronounced.

The standard error for the predicted value $\hat{y}$ at the point x is given by

$$se_{\hat{y}|x} = \hat{\sigma}\Big(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2}\Big)^{1/2}.$$

This formula is similar to (4), but there is an extra '1' inside the square root that makes $\hat{\sigma}$ a lower bound for this expression

Confidence Band for the Regression Line

A prediction interval for the value $\hat{y}$ at the point x and the $(1-\alpha)$ level is given by

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x - t_{n-2,1-\alpha/2} se_{\hat{y}|x}, \hat{\beta}_0 + \hat{\beta}_1 x + t_{n-2,1-\alpha/2} se_{\hat{y}|x}\right)$$

The `predict` function also calculates prediction intervals.

```
new.data <- data.frame(x=c(900,1100,1300))
predict(model,new.data,interval='p')
```

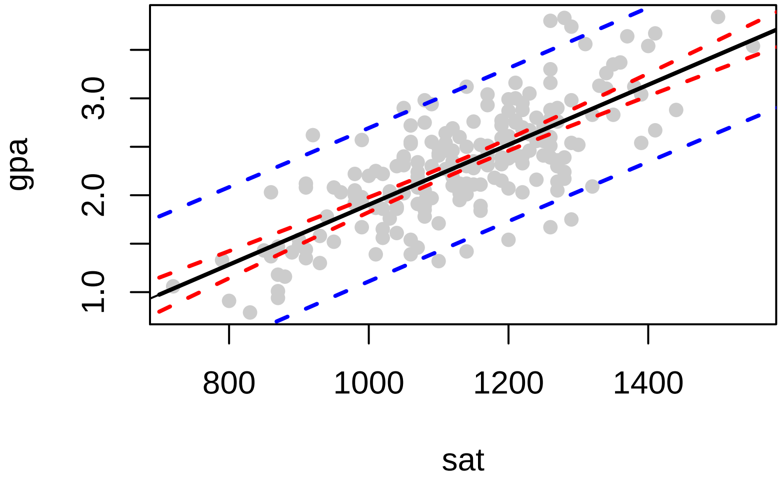
```
##        fit       lwr      upr
## 1 1.592779 0.7979927 2.387566
## 2 2.211634 1.4218455 3.001422
## 3 2.830488 2.0382696 3.622706
```

```
predict(model,new.data,interval='c')
```

```
##        fit      lwr      upr
## 1 1.592779 1.486950 1.698609
## 2 2.211634 2.154371 2.268896
## 3 2.830488 2.746088 2.914887
```

Confidence Band for the Regression Line

Let's now draw a graph including both bands for comparison.



Confidence Band for the Regression Line

```r
plot(sat, gpa)
modelA <- lm(gpa~sat, data = Grades)
abline(modelA)
new.sat <- data.frame(sat=seq(700,1600,
                              length.out = 15))
pc <- predict(modelA,new.sat, int='c')
matlines(new.sat$sat, pc, lty=c(1,2,2),lwd=rep(2,3),
         col=c('black','red','red'))
pp <- predict(modelA,new.sat, int='p')
matlines(new.sat$sat, pp, lty=c(1,2,2),lwd=rep(2,3),
         col=c('black','red','red'))
```

Confidence Band for the Regression Line

The prediction bands are much wider and include most of the observed values, as one would expect.

It is important to observe that these bands have been drawn using confidence or prediction intervals for **single values**. They are not **simultaneous** bands for the regression line.

Analysis of Variance in Linear Regression

Analysis of Variance in Linear Regression

Anova is based on dividing the sums of squares and degrees of freedom associated with the response variable Y.

The difference $y_i - \bar{y}$ is divided into two parts:

1.- The deviation of y_i from the regression line: $y_i - \hat{y}_i$.

2.- The deviation of the fitted value $\hat{y}_i$ from the mean: $\hat{y}_i - \bar{y}$.

$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$$

Squaring this relation and summing up over i

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})\end{aligned} \tag{5}$$

Analysis of Variance in Linear Regression

Let's see that the last sum in (5) is zero. Recall that $\hat{\epsilon}_i = y_i - \hat{y}_i$

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})\hat{\epsilon}_i \\ &= \sum_{i=1}^{n}\hat{y}_i\hat{\epsilon}_i - \bar{y}\sum_{i=1}^{n}\hat{\epsilon}_i\end{aligned}$$

The first sum is zero by property 3 and $\sum_i \hat{\epsilon}_i = 0$ by (1). Therefore, by (5)

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2. \tag{6}$$

Analysis of Variance in Linear Regression

This relation is commonly expressed as

$$SST = SSE + SSR$$

where

- SST denotes the total sum of squares,
- SSE is the error or residual sum of squares and
- SSR is the regression sum of squares.

Notice that the terms $y_i - \bar{y}$ represent the distance from the observed values to the average, $y_i - \hat{y}_i$ is the distance between the observed and the fitted value. In contrast, $\hat{y}_i - \bar{y}$ is the distance between the fitted value and the average observed value.

Analysis of Variance in Linear Regression

The degrees of freedom are similarly distributed.

There are $n-1$ degrees of freedom associated with SST; one degree is lost since we need to estimate the population mean μ by $\bar{y}$.

These degrees of freedom are divided into SSR and SSE.

The latter has $n-2$ degrees of freedom; two are lost because we need to calculate parameters β_0 and β_1, to fit the regression line.

Finally, there are two degrees of freedom associated with the regression line, one for the slope and one for the intercept, but one is lost since $\sum_i(\hat{y}_i - \bar{y}) = 0$ by property 1, so that SSR has one degree of freedom.

Analysis of Variance in Linear Regression

Sums of squares divided by their degrees of freedom are known as **mean squares** and are denoted by MS, thus

$$MSE = \frac{SSE}{n-2}, \qquad \text{and} \qquad MSR = \frac{SSR}{1} = SSR.$$

We have assumed that the errors in the regression are centered normal with variance σ^2, and therefore $SSE/\sigma^2 \sim \chi^2_{n-2}$, this gives $E(SSE/\sigma^2) = n-2$ and

$$E(MSE) = E\Big(\frac{SSE}{n-2}\Big) = \sigma^2,$$

which means that MSE is an unbiased estimator of σ^2.

Analysis of Variance in Linear Regression

Now let's find the expected value of MSR. Property 4 implies that $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$, therefore

$$\begin{aligned}
MSR = SSR &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{x})^2 \\
&= \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2
\end{aligned}$$

and since the values of the x_i are not random

$$\begin{aligned}
E(MSR) &= E(\hat{\beta}_1^2) \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \left(Var(\hat{\beta}_1) + \left(E(\hat{\beta}_1) \right)^2 \right) \sum_{i=1}^{n} (x_i - \bar{x})^2
\end{aligned}$$

Analysis of Variance in Linear Regression

Now, from (2) $Var(\hat{\beta}_1) = \sigma^2 v_{22}$ and by (3) $v_{22} = 1/\sum_{i=1}^{n}(x_i - \bar{x})^2$. Hence

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

On the other hand, $\hat{\beta}_1$ is an unbiassed estimator of β_1, so $\left(E(\hat{\beta}_1)\right)^2 = \beta_1^2$. Summing up

$$E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Analysis of Variance in Linear Regression

When $\beta_1 = 0$, the mean of the sampling distribution of MSR is σ^2 and coincides with the mean of MSE.

Therefore, if this hypothesis is true, the values of MSR and MSE will be similar, and the quotient MSR/MSE will be close to one.

If $\beta_1 = 0$, the quantities SSR/σ^2 and SSE/σ^2 have a χ^2 distribution with 1 and $n-2$ degrees of freedom, and it is possible to show that they are independent.

If $\beta_1 \neq 0$, then the mean of the sampling distribution of MSR will be larger than the mean of MSE by $\beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$.

Analysis of Variance in Linear Regression

In consequence,

$$\frac{MSR}{MSE} = \frac{\frac{SSR/\sigma^2}{1}}{\frac{SSE/\sigma^2}{n-2}} = \frac{\chi_1^2/1}{\chi_{n-2}^2/(n-2)} \sim F_{1,n-2}.$$

Therefore, to test $H_0 : \beta_1 = 0$ we use this statistic. If msR and msE are the observed values for the sums of squares then

$$F_{obs} = \frac{msR}{msE}$$

and large values of F_{obs} give evidence against the null hypothesis.

At a confidence level of $1 - \alpha$, the null hypothesis will be rejected if

$$F_{obs} \geq F_{1,n-2,1-\alpha}.$$

Analysis of Variance in Linear Regression

The usual way to sum up these results is through an Analysis of Variance (Anova) table.

Table 1: Anova table for example 1.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F_{obs}	Critical F
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$	`qf(1-`α`, 1, n-2)`
Error	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$		
Total	SST	$n-1$			

Analysis of Variance in Linear Regression

In `R` we get an anova table with the command `anova` acting on an object of class `lm`:

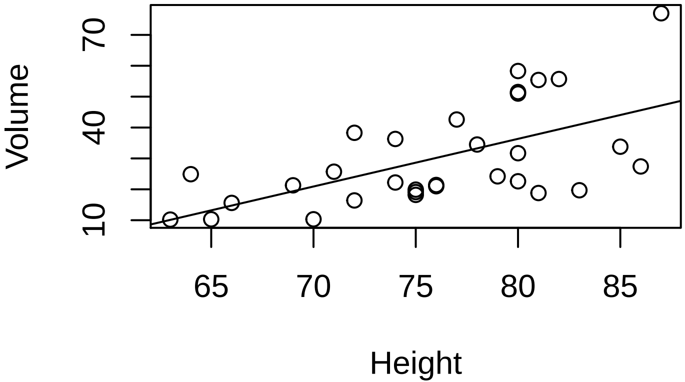
```
anova(lm1)
```

```
## Analysis of Variance Table
##
## Response: FL
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## CL          1 2329.45 2329.45  4531.1 < 2.2e-16 ***
## Residuals 198  101.79    0.51
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example 3

The data set `trees` has data on girth, height, and volume of timber in 31 felled black cherry trees. Girth is the diameter of the tree in inches measured at 4 ft 6 in above the ground.

```
plot(Volume ~ Height, data=trees)
lm4 <- lm(Volume ~ Height, data=trees)
abline(lm4)
```



Example 3

```r
summary(lm4)
```

```
##
## Call:
## lm(formula = Volume ~ Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.274  -9.894  -2.894  12.068  29.852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.1236    29.2731  -2.976 0.005835 **
## Height        1.5433     0.3839   4.021 0.000378 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 29 degrees of freedom
## Multiple R-squared:  0.3579, Adjusted R-squared:  0.3358
## F-statistic: 16.16 on 1 and 29 DF,  p-value: 0.0003784
```

Example 3

```r
anova(lm4)
```

```
## Analysis of Variance Table
##
## Response: Volume
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Height     1 2901.2 2901.19  16.165 0.0003784 ***
## Residuals 29 5204.9  179.48
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

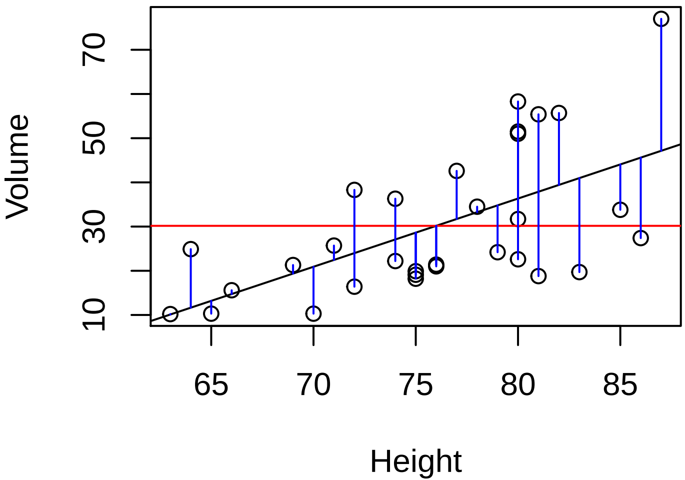
Diagnostic plots

Two useful functions for extracting information from an object of class `lm` are `resid` and `fitted`, that will give the residuals and fitted values, respectively.

Let us use them in this example to graph the residuals in the previous plot

```
plot(Volume ~ Height, data=trees)
abline(lm4)
abline(h=mean(trees$Volume), col='red')
segments(trees$Height,fitted(lm4),
         trees$Height,trees$Volume, col='blue')
```

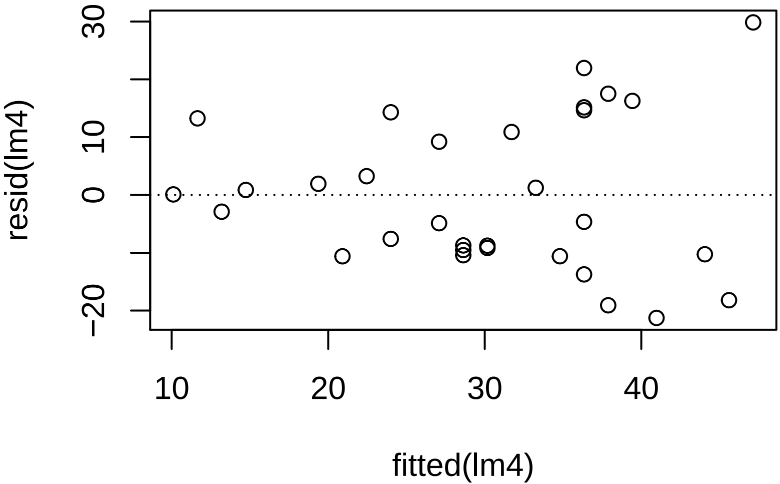
Diagnostic plots



Diagnostic plots

We can also plot the fitted values versus residuals

```
plot(fitted(lm4),resid(lm4))
abline(h=0, lty='dotted')
```

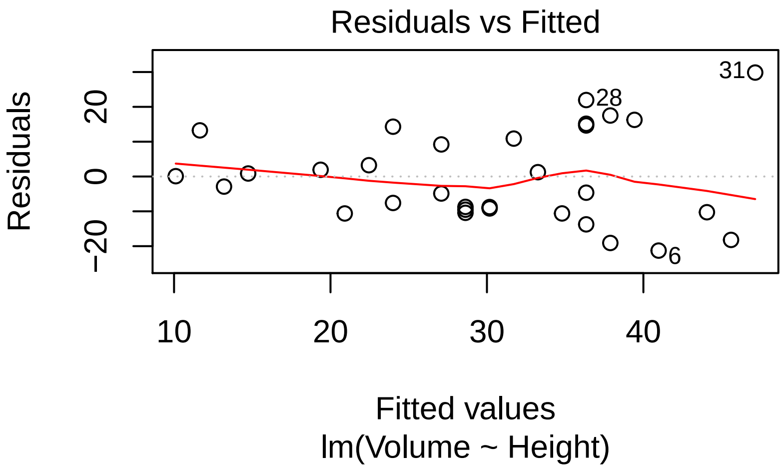


Diagnostic plots

This plot is part of the diagnostic plots that are usually made to evaluate the goodness of fit of the model and the validity of the assumption. They can be obtained with the following instructions.

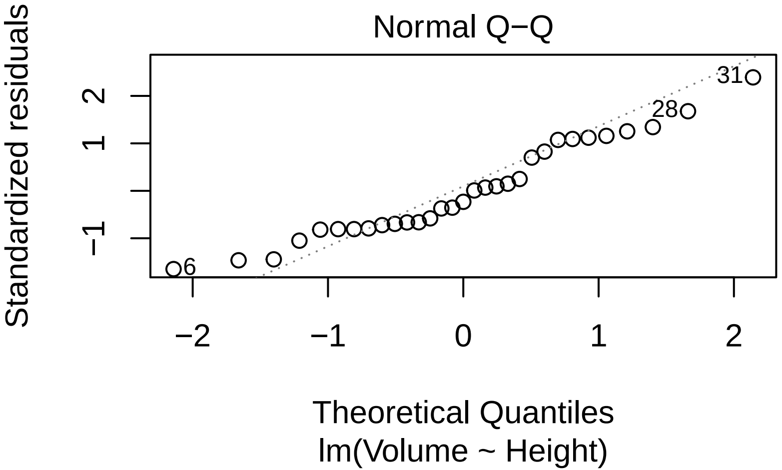
Diagnostic plots

```r
plot(lm4, which = 1)
```



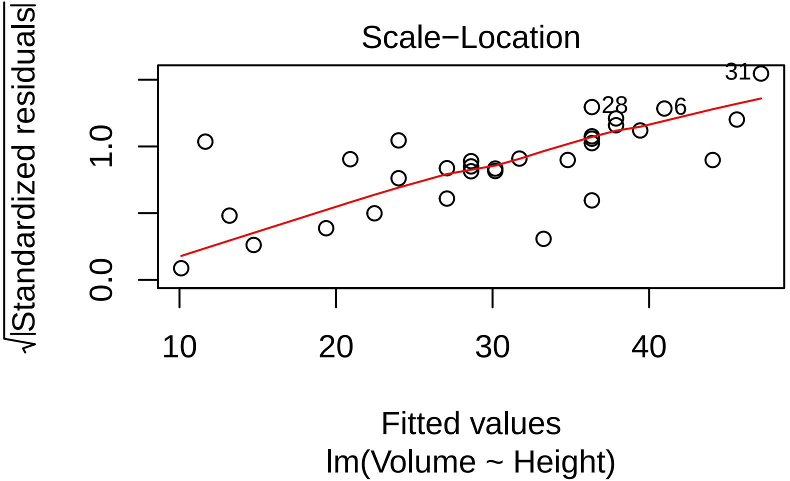
Diagnostic plots

```r
plot(lm4, which = 2)
```



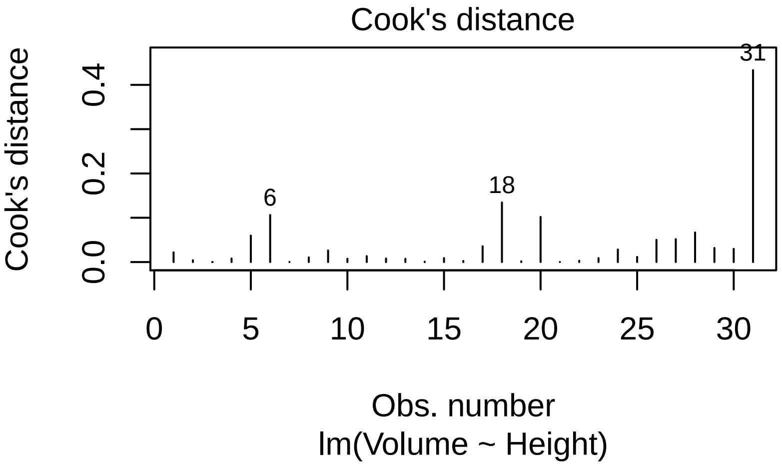
Diagnostic plots

```r
plot(lm4, which = 3)
```



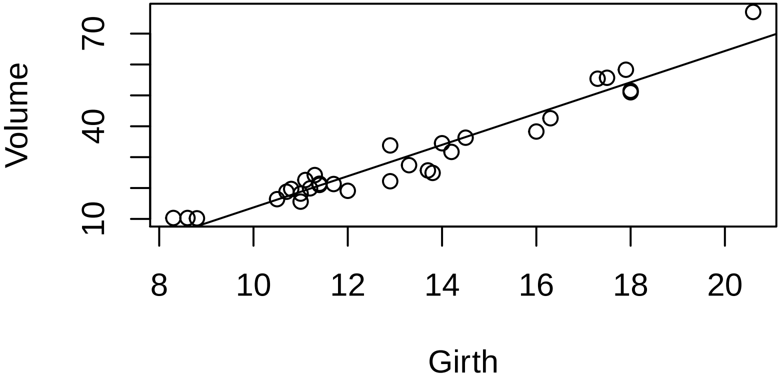
Diagnostic plots

```r
plot(lm4, which = 4)
```



Example 3: Another model

```r
plot(Volume~Girth, data=trees)
lm5 <- lm(Volume ~Girth, data = trees)
abline(lm5)
```



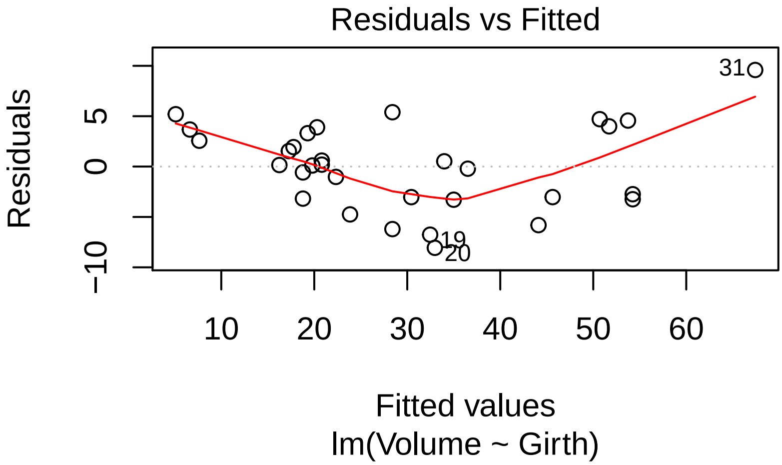
Example 3: Another model

```r
summary(lm5)
```

```
##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.065 -3.107  0.152  3.495  9.587
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
## Girth         5.0659     0.2474   20.48  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 29 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
## F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

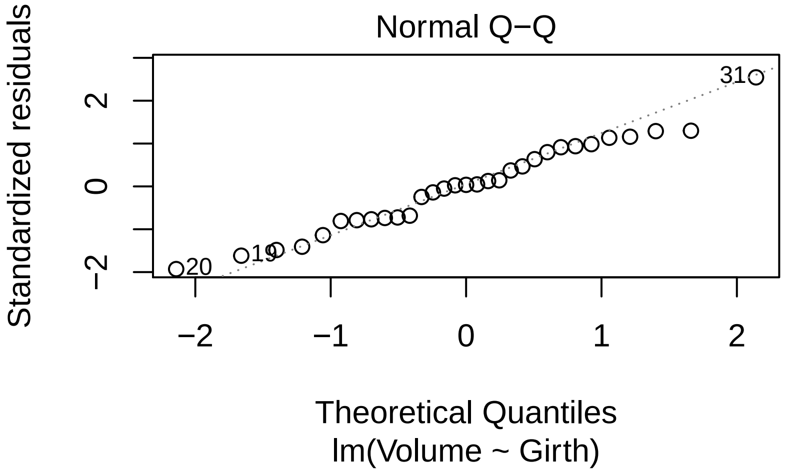
Diagnostic plots

```
plot(lm5, which = 1)
```



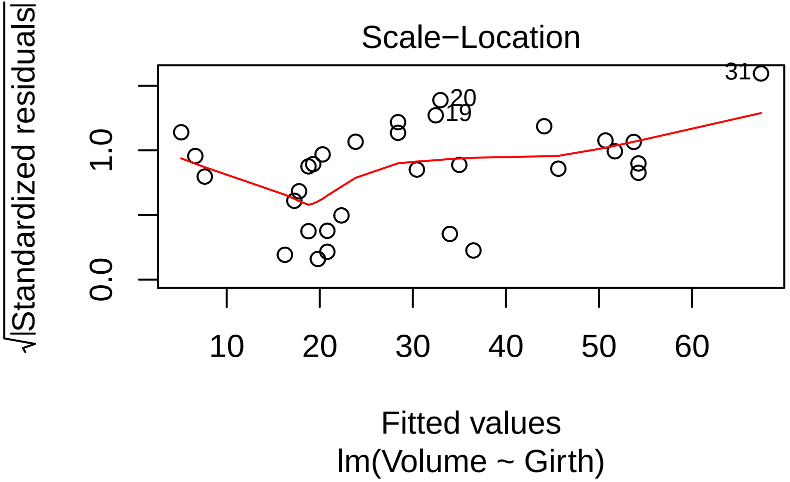
Diagnostic plots

```r
plot(lm5, which = 2)
```



Diagnostic plots

```r
plot(lm5, which = 3)
```



Diagnostic plots

```
plot(lm5, which = 4)
```