Linear Regression Analysis

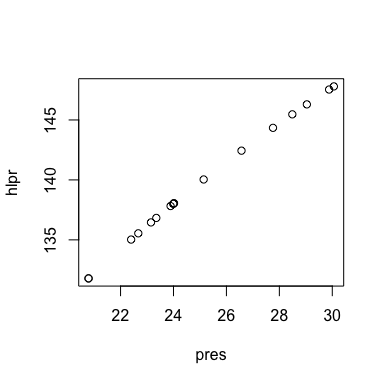
### Example: Association of boiling point and pressure

In the 1840s and 1850s a Scottish physicist, James D. Forbes, wanted to be able to estimate altitude above sea level from measurement of the boiling point of the water. He studied the relationship between boiling point and pressure. Forbes’ theory suggested that over the range of observed values the graph of boiling point versus the logarithm of pressure yields a straight line. Since the logs of the pressures do not vary much, all values of log(press) are multiplied by 100. This avoids studying very small numbers, without changing the major features of the analysis.  
 *Reference: S. Weisberg. Applied Linear Regression. Wiley 2005.*

# load the data and attach the dataset  
library(MASS)  
data(forbes)  
  
# Transform the pressure to 100\*log(pressure) using a base 10 logarithm  
forbes$hlpr<-100\*log10(forbes$pres)  
forbes$boil<-forbes$bp

Create a scatter plot of the predictor versus the response.

plot(hlpr~pres, data=forbes)



**This looks like a reasonable linear relationship.**

Create summary statistics and calculate the correlation coefficient

summary(forbes$boil)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 194.3 199.4 201.3 203.0 208.6 212.2

summary(forbes$hlpr)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 131.8 136.5 138.0 139.6 144.3 147.8

cor(forbes$boil,forbes$hlpr)

## [1] 0.9974771

Fit a linear regression model and look at the regression output

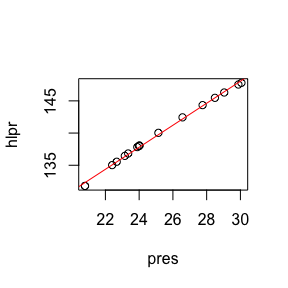
fit<-lm(hlpr~boil, data=forbes)  
summary(fit)

##   
## Call:  
## lm(formula = hlpr ~ boil, data = forbes)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.31974 -0.14707 -0.06890 0.01877 1.35994   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -42.16418 3.34136 -12.62 2.17e-09 \*\*\*  
## boil 0.89562 0.01646 54.42 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3792 on 15 degrees of freedom  
## Multiple R-squared: 0.995, Adjusted R-squared: 0.9946   
## F-statistic: 2962 on 1 and 15 DF, p-value: < 2.2e-16

**The boiling point is a sigificant predictor for 100 log(pressure).**

Draw a scatter plot with the regression line.

plot(hlpr~pres, data=forbes)  
abline(lm(hlpr~pres, data=forbes), col="red")



Regression coefficients: this is the slope

fit$coefficients[2]

## boil   
## 0.8956178

Note that the estimated regression line is hlpr=fit$coefficients[1]+fit$coefficients[2]\*boil namely hlpr= -42.1641838 + 0.8956178 boil.

R square and adjusted R square indicate how much variation of the response (hlpr) is explained by regression on the boiling point.

summary(fit)$r.squared

## [1] 0.9949606

summary(fit)$adj.r.squared

## [1] 0.9946246

How scattered are the points around the line?

summary(fit)$sigma

## [1] 0.3791592

Calculate a 95% CI for the slope

tble<-summary(fit)$coefficients  
slope.mean<-tble[2,1]  
slope.se<-tble[2,2]  
df<-fit$df.residual  
critval<-abs(qt(0.025,df))  
  
CIlow<-slope.mean-critval\*slope.se  
CIup<-slope.mean+critval\*slope.se

A 95% CI for the slope is (0.8605395, 0.9306961).

Here are a number of summaries:

coefficients(fit) # model coefficients

## (Intercept) boil   
## -42.1641838 0.8956178

confint(fit, level=0.95) # CIs for model parameters

## 2.5 % 97.5 %  
## (Intercept) -49.2861237 -35.0422438  
## boil 0.8605395 0.9306961

fitted(fit) # predicted values

## 1 2 3 4 5 6 7 8   
## 132.0335 131.8543 135.0786 135.5264 136.4220 136.8698 137.7654 137.9445   
## 9 10 11 12 13 14 15 16   
## 138.2132 138.1237 140.1836 141.0792 145.4677 144.6617 146.5425 147.6172   
## 17   
## 147.8859

residuals(fit) # residuals

## 1 2 3 4 5 6   
## -0.24802254 -0.06889899 -0.05377004 0.01877126 0.03310101 -0.04111891   
## 7 8 9 10 11 12   
## 0.05618981 0.05847608 -0.15593374 -0.08445627 -0.14706580 1.35994454   
## 13 14 15 16 17   
## 0.00150698 -0.31973578 -0.24281806 -0.07916126 -0.08700828

anova(fit) # anova table

## Analysis of Variance Table  
##   
## Response: hlpr  
## Df Sum Sq Mean Sq F value Pr(>F)   
## boil 1 425.76 425.76 2961.5 < 2.2e-16 \*\*\*  
## Residuals 15 2.16 0.14   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

influence(fit) # regression diagnostics

## $hat  
## 1 2 3 4 5 6   
## 0.19344032 0.19988585 0.10692651 0.09787772 0.08260614 0.07638336   
## 7 8 9 10 11 12   
## 0.06676382 0.06529208 0.06336706 0.06397105 0.05961234 0.06393448   
## 13 14 15 16 17   
## 0.13957975 0.11890328 0.17189609 0.20963838 0.21992176   
##   
## $coefficients  
## (Intercept) boil  
## 1 -1.011985595 4.897179e-03  
## 2 -0.289972641 1.403809e-03  
## 3 -0.119867502 5.731667e-04  
## 4 0.037448196 -1.784857e-04  
## 5 0.051140082 -2.415222e-04  
## 6 -0.054588130 2.560660e-04  
## 7 0.050804751 -2.328767e-04  
## 8 0.048004441 -2.183974e-04  
## 9 -0.108649482 4.870900e-04  
## 10 -0.062334329 2.809853e-04  
## 11 0.029493232 -1.906479e-04  
## 12 -0.829501124 4.508246e-03  
## 13 -0.004281493 2.160362e-05  
## 14 0.762207722 -3.860766e-03  
## 15 0.851335765 -4.279732e-03  
## 16 0.336754638 -1.688304e-03  
## 17 0.387809910 -1.943165e-03  
##   
## $sigma  
## 1 2 3 4 5 6 7   
## 0.3854642 0.3919268 0.3921724 0.3924315 0.3923584 0.3923004 0.3921591   
## 8 9 10 11 12 13 14   
## 0.3921340 0.3900975 0.3917730 0.3903685 0.1135960 0.3924668 0.3817627   
## 15 16 17   
## 0.3859336 0.3917449 0.3915829   
##   
## $wt.res  
## 1 2 3 4 5 6   
## -0.24802254 -0.06889899 -0.05377004 0.01877126 0.03310101 -0.04111891   
## 7 8 9 10 11 12   
## 0.05618981 0.05847608 -0.15593374 -0.08445627 -0.14706580 1.35994454   
## 13 14 15 16 17   
## 0.00150698 -0.31973578 -0.24281806 -0.07916126 -0.08700828

Is the boiling point statistically significantly associated with HLPR?

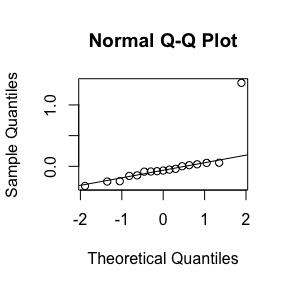
pval<-tble[2,4]

**The p value for testing vs is 1.189783210^{-18}.**

### Diagnostics and Residual Analysis

Is there evidence that the residuals do not follow a normal distribution?

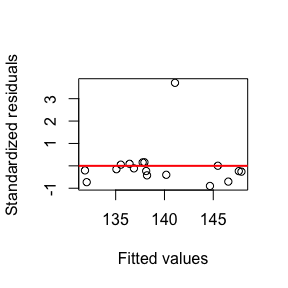
qqnorm(fit$residuals)  
qqline(fit$residuals)



**It appears there is one outlier.**

Make a residual plot.

fit.stdres = rstandard(fit)  
plot(fit$fitted, fit.stdres, xlab="Fitted values", ylab="Standardized residuals")  
abline(h=0, col=2, lwd=2)



You can identify the point with the largest residual by using identify(fit$fitted, fit.stdres) click on the point in questions and then click Esc. This will identify the point as the 12-th record in the dataset.

In the presence of outliers, it is useful to do the analysis with and without the point. Remove large residuals.

indx<-which(fit.stdres>2)  
forbes1<-forbes[-indx,]

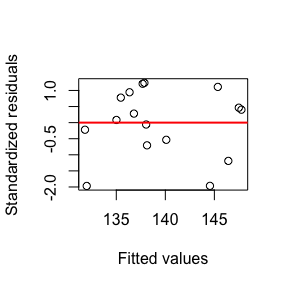
Redo regression analysis with this outlier removed

fit1<-lm(hlpr~boil, data=forbes1)  
summary(fit1)

##   
## Call:  
## lm(formula = hlpr ~ boil, data = forbes1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.20882 -0.06338 0.01974 0.08842 0.13558   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -41.334683 1.003312 -41.2 5.16e-16 \*\*\*  
## boil 0.891110 0.004944 180.2 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1136 on 14 degrees of freedom  
## Multiple R-squared: 0.9996, Adjusted R-squared: 0.9995   
## F-statistic: 3.249e+04 on 1 and 14 DF, p-value: < 2.2e-16

**The regression fit and coefficient of determination did not change much.**

fit.stdres = rstandard(fit1)  
plot(fit1$fitted, fit.stdres, xlab="Fitted values", ylab="Standardized residuals")  
abline(h=0, col=2, lwd=2)



**There are no obvious outliers in this plot after the point was removed.**