Chapter 2

Linear Regression

The normal linear regression takes the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i. \tag{2.1}$$

- y_i is the ith observed value of the *response variable*, $i = 1, \cdots, n$
- x_{ki} is the ith observed value of the *explanatory variable* x_k, $k = 1, 2, \cdots, p$
- ε_i is the *error* or *residual* for each individual i, and is assumed to have a normal distribution with mean 0 and variance σ^2
- β_0 is the intercept parameter
- β_k $(k = 1, 2, \cdots, p)$ is the predicted change in the response variable when the explanatory variable x_k increases by one unit. The sign of β_k indicates a positive or negative relationship. $\beta_k = 0$ indicates the absence of any linear relationship between the explanatory variable x_k and the response y.

When $p = 1$, it is a simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{2.2}$$

When $p > 1$, it is called a multiple linear regression.

The response and explanatory variables can be

- categorical
  - two categories, also known as binary
  - more than two categories, unordered
  - ordered
- discrete counts
- continuous

Estimation of the parameters

The aim is to determine the values of the parameters in the model such that the model is the best fit to the data. This is done by using the *method of least squares* for the normal linear regression. The construction of the least squares estimates $\hat{\beta}_k$ $(k = 0, 1, 2, \cdots, p)$ do not require any assumption about the *errors* ε_i.

If we can assume that the errors are independent, normally distributed, and have equal variance, then the maximum likelihood estimates of the parameters are identical with the least squares estimates.

The *predicted* or *fitted* values are

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_p x_{pi}.$$

The *residuals* are $y - \hat{y}$.

Model checking

A useful method to examine the fit of a linear regression is to check the residuals $y - \hat{y}$.

The following three plots are helpful to check for systematic departure of the model from the data:

- the residuals against an explanatory variable (check curvature)
- the residuals against the fitted values (check heteroscedasticity)
- a normal probability plot of the residuals (check non-normality of errors)

qqplot. Compare against normal dist.

Outliers and influential observations

The above three plots may reveal *outliers* in the data. A regression model can be strongly influenced by an outlier. It is therefore important to check whether an observation is indeed an outlier or appears to be an outlier because of misspecification of the model.

It is possible that an outlier may have little *influence* on the fitted model. In this case, we do not need to worry about it.

It is also possible that an observation has a substantial *influence* on the fitted model but is not an outlier. The *leverage* h_i of observation i is an overall measure of the potential influence that observation can have on the analysis. The quantity h_i is the ith diagonal element of the "hat" matrix. For a normal linear model, it is a measure of the distance of that observation from the mean of the explanatory variable. In general, we should examine observations which have $h_i > 2m/n$ to see if they highly affect the values and precision of the parameter estimates, where m is the number of parameters.

A direct measure of the influence of an observation on the analysis is *Cook's distance* C_i. This combines leverage and residuals in a single measure and summarizes the influence on the parameter estimates if one observation is removed from the data set. An observation with $C_i > 8/(n-2m)$ deserves a closer look.

Model selection

For two nested models

- Model 1: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_q x_{qi} + \varepsilon_i$
- Model 2: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_q x_{qi} + \cdots + \beta_p x_{pi} + \varepsilon_i$ with $q < p$

we can use the *analysis of variance* to compare them and thus select the "best" model.

The `step` command in R has an option to allow the selection of the model to involve *forward* selection, *backward* selection or both (*stepwise* selection). The default is backwards elimination if the `scope` argument is missing.

- *Forward* selection introduces new variables one at a time, and stops with the best model.
- *Backward* selection starts with all the variables available and drops the least significant variable one at a time, and stops when no variable can be deleted without increasing the sum of squares significantly.
- *Stepwise* selection allows four options at each step: adding a variable, deleting a variable, swapping a variable in the model for one not in the model, or stopping.

Alternatively, a popular method that is adopted by many statisticians to determine a suitable model is based on a model selection criterion, such as the Akaike's Information Criterion (AIC). This method does not require the models to be nested. The Akaike's Information Criterion is defined as

$$AIC = -2l + 2m$$

where l is the maximized log-likelihood for the model and m is the number of parameters in the model. We choose the model that has the smallest AIC. For a general linear model,

$$AIC = n \ln(RSS/n) + 2m$$

where ln is the natural log, and RSS is the *residual sum of squares.*

You can get different AIC values when using different R commands. This is potentially confusing: it arises because the maximized log-likelihood involves a term that can be ignored as it is the same whatever model we fit ("constant term"). In some situations this constant term will be calculated, and in others it won't. The key point is that *the difference in the AIC values for two models will not be affected.*

2.1 Simple linear regression

Example

The Janka hardness test measures the force required to push a steel ball with a diameter of 11.28 millimeters into the wood to a depth of half the ball's diameter. It is not easy to measure Janka hardness directly. The density of the wood can be used to infer about the hardness.

Data were collected on the density (`denst`) and Janka hardness (`hard`) of 36 Australian eucalypt hardwoods to help establish relationship between the two variables (Williams, 1959, *Regression analysis.* John Wiley & Sons Inc., New York).

```
> eucalypt <- read.table("eucalypt.txt",header=T)
> attach(eucalypt)
> names(eucalypt)
```

```
[1] "denst" "hard"
```

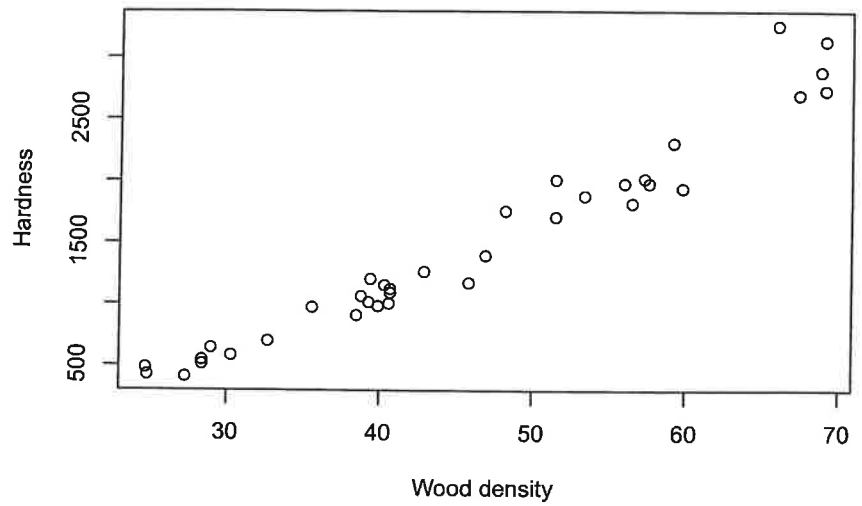
```
> head(eucalypt)
```

```
  denst hard
1  24.7  484
2  24.8  427
3  27.3  413
4  28.4  517
5  28.4  549
6  29.0  648
```

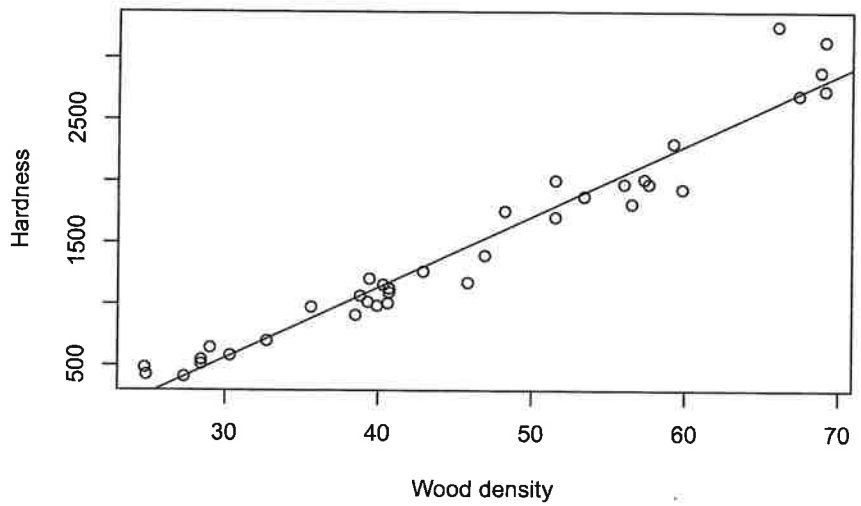
```
> tail(eucalypt)
```

```
   denst hard
31  59.8 1940
32  66.0 3260
33  67.4 2700
34  68.8 2890
35  69.1 2740
36  69.1 3140
```

```
> plot(denst,hard,xlab="Wood density",ylab="Hardness")
```




```
> a <- lm(hard~denst)
> plot(denst,hard,xlab="Wood density",ylab="Hardness")
> abline(a)
```




```
> summary(a)
```

```
.....
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1160.500    108.580  -10.69 2.07e-12 ***
denst          57.507      2.279   25.24  < 2e-16 ***

Residual standard error: 183.1 on 34 degrees of freedom
Multiple R-squared:  0.9493,        Adjusted R-squared:  0.9478
F-statistic:   637 on 1 and 34 DF,  p-value: < 2.2e-16
```

testing whether parameters are equal to 0.

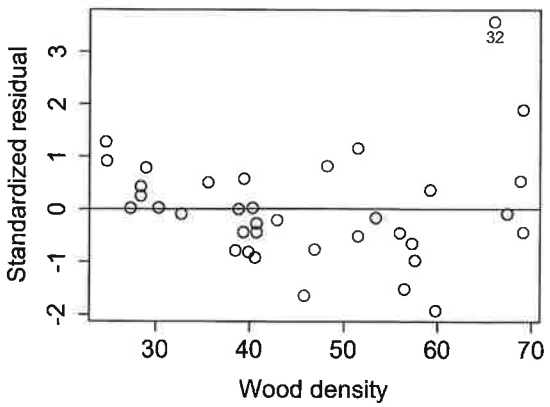
The fitted model is $\hat{y} = -1160.500 + 57.507x$. The 95% confidence intervals for the estimated parameters are obtained by

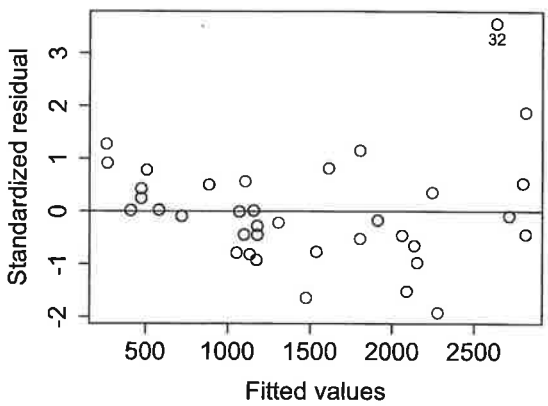
```
> confint(a)
                  2.5 %     97.5 %
(Intercept) -1381.16001 -939.83940
denst          52.87614   62.13721
```

```
> anova(a)
Analysis of Variance Table

Response: hard
          Df   Sum Sq  Mean Sq F value    Pr(>F)
denst      1 21345674 21345674  636.98 < 2.2e-16 ***
Residuals 34  1139366    33511
```

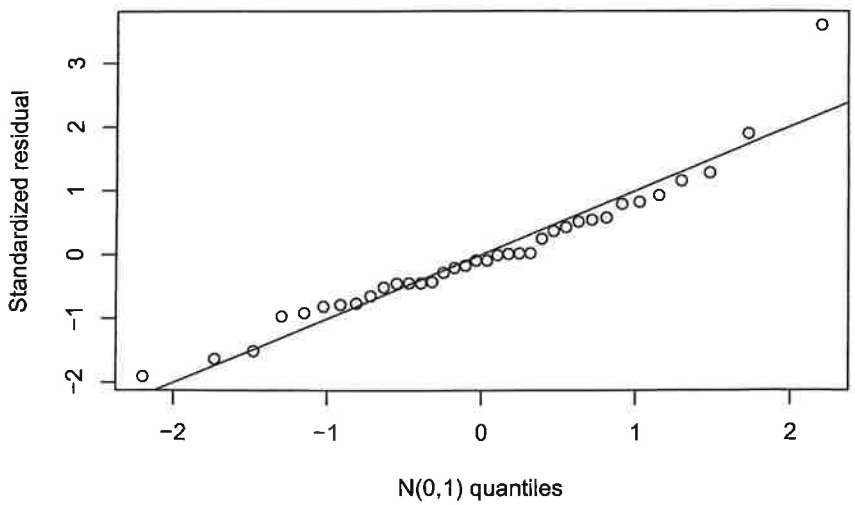
We plot the standardized residuals against both the explanatory variable and the fitted values.

```
> r = rstandard(a)
> lf <- fitted(a)
> plot(denst,r,xlab="Wood density",ylab="Standardized residual")
> abline(h=0)
> plot(lf,r,xlab="Fitted values",ylab="Standardized residual")
> abline(h=0)
```





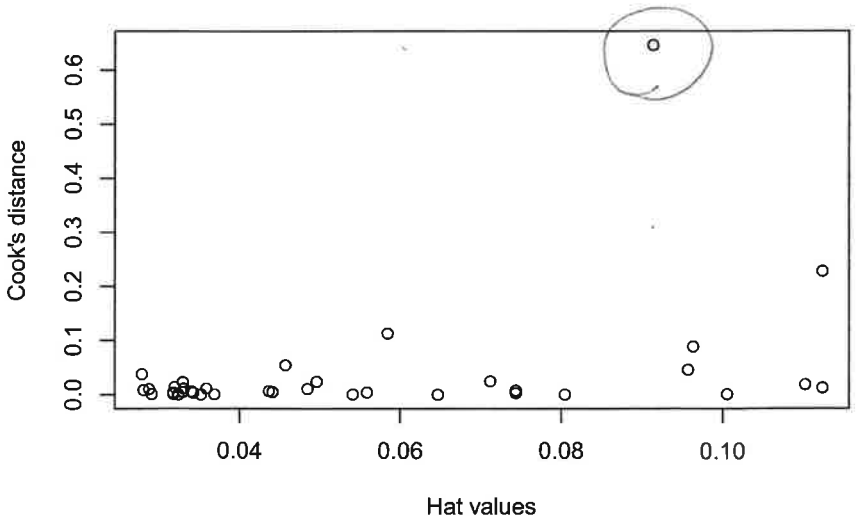
Now we check the normality of the residuals.




```
> qqnorm(r,xlab="N(0,1) quantiles",ylab="Standardized residual")
> abline(0,1)
```

It seems that there is one influential point.

```
> h <- hatvalues(a)
> cd <- cooks.distance(a)
> plot(h,cd,xlab="Hat values",ylab="Cook's distance")
```



To account for the curvature in the residuals, we will next examine if a quadratic term of the explanatory variable will improve the model fit.

```
> denst2 <- denst^2
> b <- lm(hard~denst+denst2)
> summary(b)
```

```
lm(formula = hard ~ denst + denst2)

......
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -118.0074   334.9669  -0.352  0.72686
denst          9.4340    14.9356   0.632  0.53197
denst2         0.5091     0.1567   3.248  0.00267 **

Residual standard error: 161.7 on 33 degrees of freedom
Multiple R-squared:  0.9616,    Adjusted R-squared:  0.9593
F-statistic: 413.2 on 2 and 33 DF,  p-value: < 2.2e-16
```

```
> c <- lm(hard~denst2)
> summary(c)
```

```
lm(formula = hard ~ denst2)

......
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 90.69861   54.53844   1.663    0.106
denst2       0.60717    0.02094  28.999   <2e-16 ***

Residual standard error: 160.3 on 34 degrees of freedom
Multiple R-squared:  0.9611,    Adjusted R-squared:   0.96
F-statistic: 840.9 on 1 and 34 DF,  p-value: < 2.2e-16
```

```
> anova(b,c)
```

Test whether necessary

```
Analysis of Variance Table

Model 1: hard ~ denst + denst2
Model 2: hard ~ denst2
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     33 863325
2     34 873763 -1    -10438 0.399  0.532
```

nested

2 models are not significantly different.

∴ Choose simpler model so model "c"

```
> AIC(a); AIC(b); AIC(c)
```

```
[1] 481.2123
[1] 473.2246
[1] 471.6572
```

```
> AIC(a)-AIC(b)
```

```
[1] 7.987695
```

```
> AIC(b)-AIC(c)
```

```
[1] 1.567363
```

```
> step(b)
```

```
Start:  AIC=369.06
hard ~ denst + denst2

         Df Sum of Sq     RSS    AIC
- denst   1     10438  873763 367.49
<none>                 863325 369.06
- denst2  1    276041 1139366 377.05

Step:  AIC=367.49
hard ~ denst2

         Df Sum of Sq      RSS    AIC
<none>                 873763 367.49
- denst2  1  21611278 22485041 482.41

Call:
lm(formula = hard ~ denst2)

Coefficients:
(Intercept)       denst2
    90.6986       0.6072
```

which vars are dropped

don't drop anything

```
> 377.05-369.06
```

```
[1] 7.99
```

```
> 369.06-367.49
```

```
[1] 1.57
```

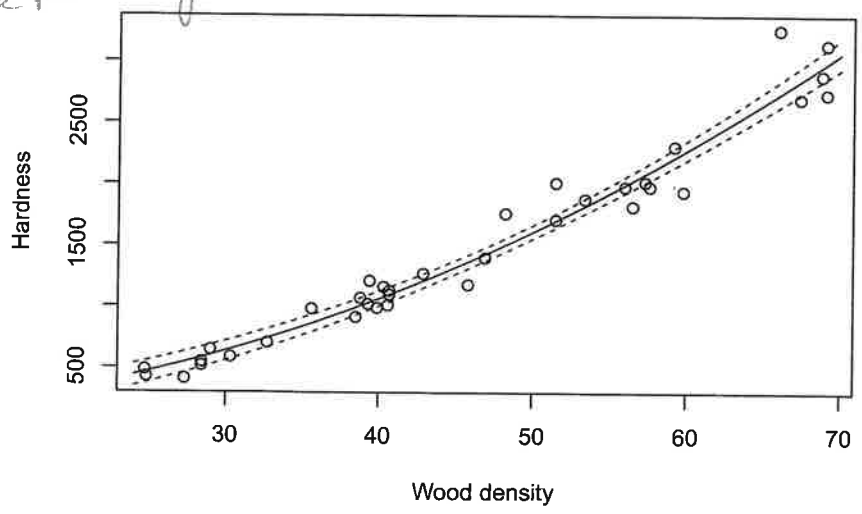
The fitted model c is $\hat{y} = 90.699 + 0.607x^2$. The 95% confidence intervals for the estimated parameters are obtained by

```
> confint(c)
```

```
                2.5 %      97.5 %
(Intercept) -20.1368238 201.5340513
denst2        0.5646161   0.6497163
```

We use model c to predict the fitted vales for a smooth range of `denst` values between 24 and 70. Then we can plot the linear regression with the confidence intervals of the predicted values overlaid.

```
> plot(denst,hard,xlab="Wood density",ylab="Hardness")
> x <- seq(from=24,to=70,by=1)
> x2 <- x^2
> y <- predict(c,list(denst2=x2),interval="confidence")
> matlines(x,y,lty=c(1,2,2),col="black")
```

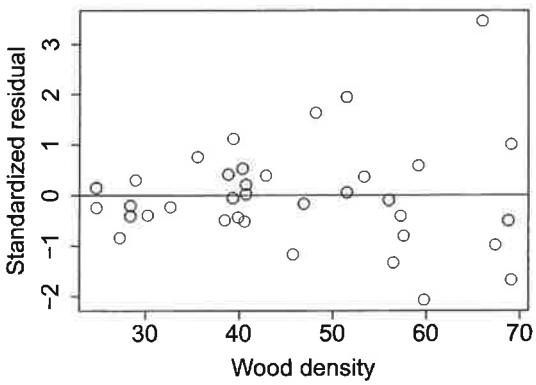


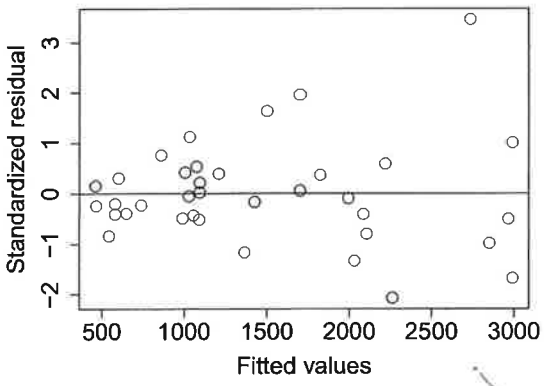
Now, we can ask for the predicted hardness with confidence intervals if the wood density was 57. This can be done by

```
> x2 <- 57^2
> predict(c,list(denst2=x2),interval="confidence")
```

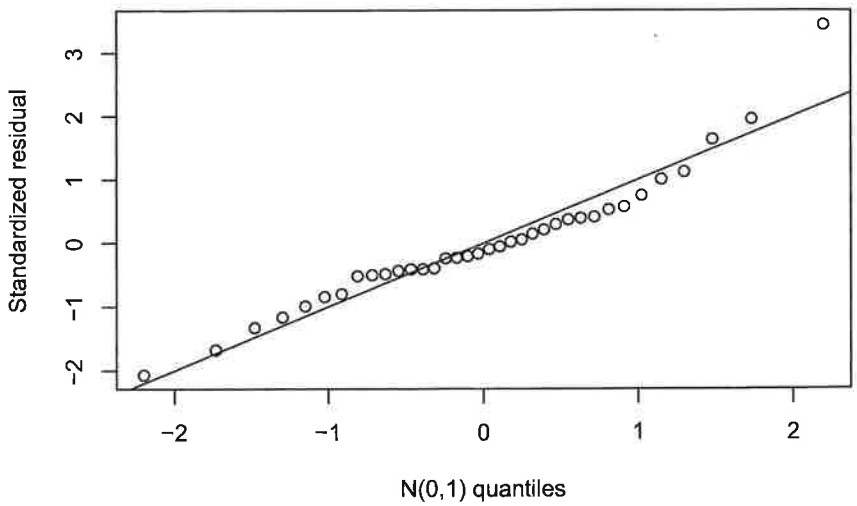
```
       fit      lwr      upr
1 2063.382 1994.967 2131.796
```

```
> r <- rstandard(c)
> lf <- fitted(c)
> plot(denst,r,xlab="Wood density",ylab="Standardized residual")
> abline(h=0)
> plot(lf,r,xlab="Fitted values",ylab="Standardized residual")
> abline(h=0)
```



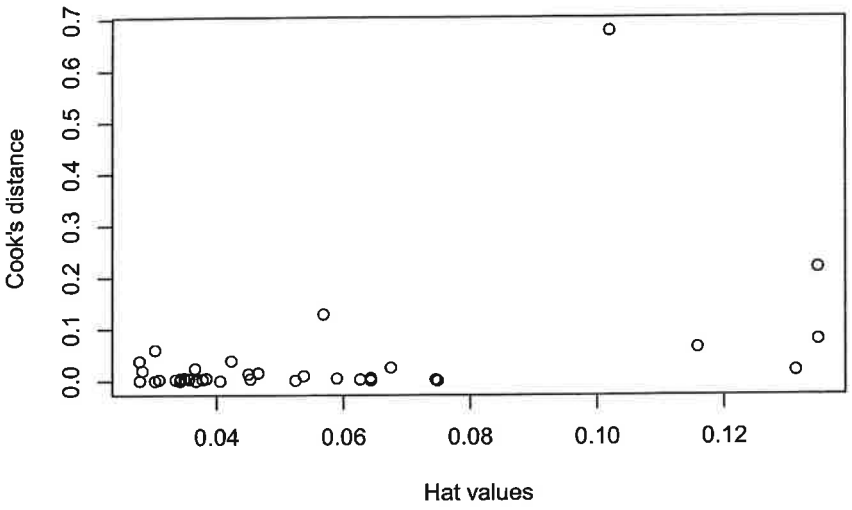



```
> qqnorm(r,xlab="N(0,1) quantiles",ylab="Standardized residual")
> abline(0,1)
```



increasing variance with the mean.

```
> h <- hatvalues(c)
> cd <- cooks.distance(c)
> plot(h,cd,xlab="Hat values",ylab="Cook's distance")
```



Transforming the response

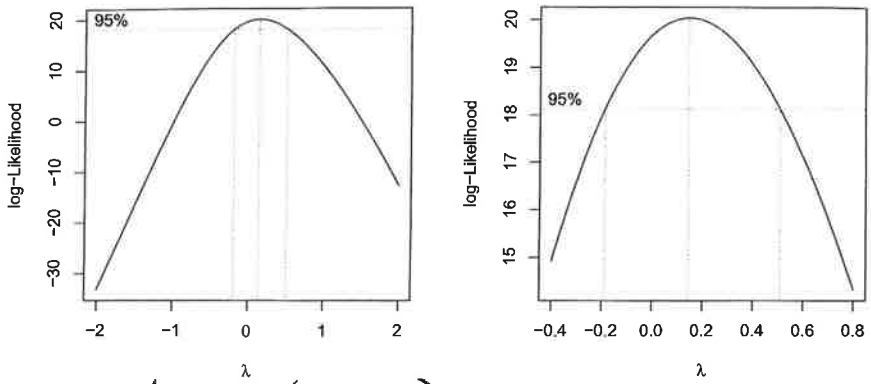
It seems that the variance is increasing with the mean. We can use the R function `bptest()` in the package `lmtest` to test whether the estimated variance of the residuals from the above best regression model `c` is dependent on the values of the explanatory variable.

```
> library(lmtest)
> bptest(hard~denst2)

        studentized Breusch-Pagan test

data:  hard ~ denst2
BP = 6.7526, df = 1, p-value = 0.009361
```

null hypot. not supported

⇒ variance IS dependent on explanatory variable

We can use Box-Cox transformation to look for an optimal transformation of the response variable. To use Box-Cox transformation, the response variable has to be positive. The value of λ (lambda) that maximizes the likelihood when a specified set of explanatory variables is used is the suggested power transformation of the response variable. When $\lambda = -1$, it suggests an inverse transformation of the response variable; $\lambda = 1$ suggests the original scale; $\lambda = 0.5$ suggests the square-root transformation; and $\lambda = 0$ suggests the logarithm transformation.



library (MASS)

```
> boxcox(hard~denst+denst2)
> boxcox(hard~denst+denst2,lambda=seq(-0.4,0.8,0.1))
```

From the Box-Cox analysis, the most obvious transformation would be the logarithm given that $\lambda = 0$ is in the 95% confidence interval.

```
> a2 <- lm(log(hard)~denst+denst2)
> summary(a2)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.138e+00  2.087e-01  19.828  < 2e-16 ***
denst        9.152e-02  9.305e-03   9.835 2.45e-11 ***
denst2      -5.228e-04  9.764e-05  -5.354 6.49e-06 ***

Residual standard error: 0.1008 on 33 degrees of freedom
Multiple R-squared:  0.9723,    Adjusted R-squared:  0.9706
F-statistic: 578.9 on 2 and 33 DF,  p-value: < 2.2e-16
```

```
> anova(a2)
```

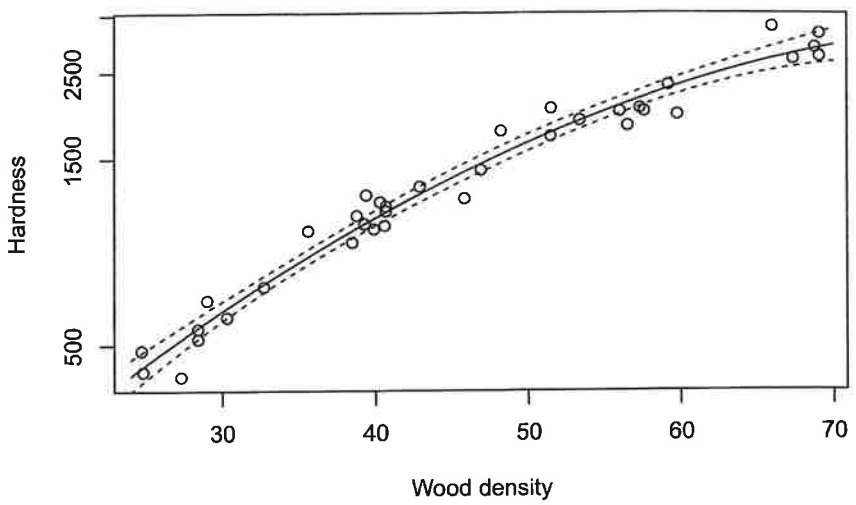
```
Analysis of Variance Table

Response: log(hard)
          Df  Sum Sq Mean Sq  F value    Pr(>F)
denst      1 11.4665 11.4665 1129.184 < 2.2e-16 ***
denst2     1  0.2911  0.2911   28.668 6.486e-06 ***
Residuals 33  0.3351  0.0102
```

both reduce significantly the Sum of Squares error

```
> plot(denst,log(hard),xlab="Wood density",ylab="Hardness",axes=F)
> axis(1)
> axis(2,at=log(seq(500,3500,1000)),seq(500,3500,1000))
> box()
```

plotting with original (not log transformed) axis

```
> x <- seq(from=24,to=70,by=1)
> x2 <- x^2
> y <- predict(a2,list(denst=x,denst2=x2),interval="confidence")
> matlines(x,y,lty=c(1,2,2),col="black")
```



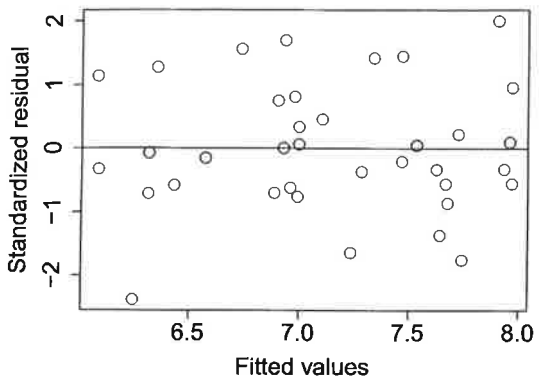
The predicted *median* hardness and confidence interval of the *median* hardness for a wood density of 57 can be obtained by

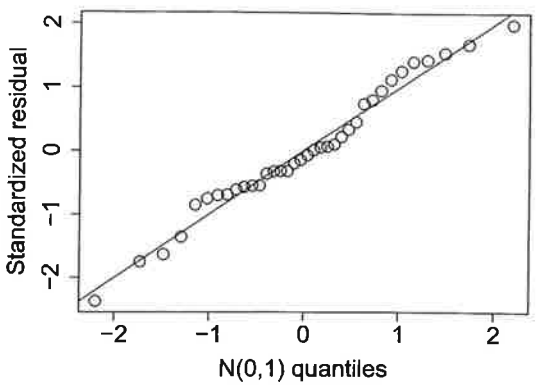
```
> xp <- 57
> xp2 <- xp^2
> pred <- predict(a2,list(denst=xp,denst2=xp2),interval="confidence")
> exp(pred)
```

```
       fit      lwr      upr
1 2112.995 2014.379 2216.439
```

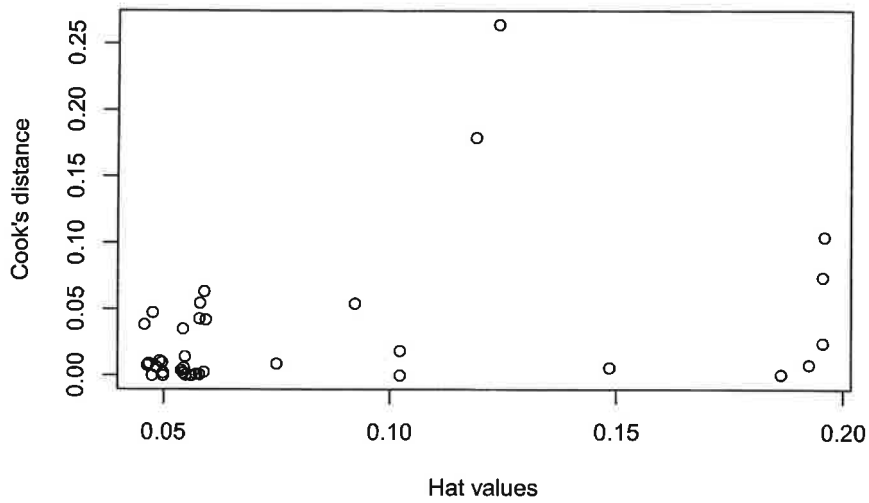
```
> r = rstandard(a2)
> lf <- fitted(a2)
> plot(lf,r,xlab="Fitted values",ylab="Standardized residual")
> abline(h=0)
> qqnorm(r,xlab="N(0,1) quantiles",ylab="Standardized residual")
> abline(0,1)
```

predicted median hardness when you transform back






```
> h <- hatvalues(a2)
> cd <- cooks.distance(a2)
> plot(h,cd,xlab="Hat values",ylab="Cook's distance")
```



Test if variance still increasing?

```
> bptest(log(hard)~denst+denst2)

        studentized Breusch-Pagan test

data:  log(hard) ~ denst + denst2
BP = 0.1734, df = 2, p-value = 0.917
```

NO!

Null hypothesis - variance dependent on explanatory variable.

Note that the model with the logarithm transformation of the response variable fits the data the best. However, whether this is the right relationship between the wood density and Janka hardness needs consultation with experts in this applied field. Sometimes we may need to find other explanation (such as extra factors which may affect the response variable) of the non-constant variance.

2.2 Multiple linear regression

Example

Woods et al. (1932, Industrial Engineering and Chemistry, 24, 1207-1214) investigated the relationship between the heat evolved during the setting of cement and its compounds composition, namely, tricalcium aluminate ($3CaO \cdot Al_2O_3$), tricalcium silicate ($3CaO \cdot SiO_2$), tetracalcium aluminoferrite ($4CaO \cdot Al_2O_3 \cdot Fe_2O_3$), and β-dicalcium silicate ($2CaO \cdot SiO_2$). We look at the 13 samples listed in Table 8.1 in Davison (2003, *Statistical Models*, Cambridge University Press).

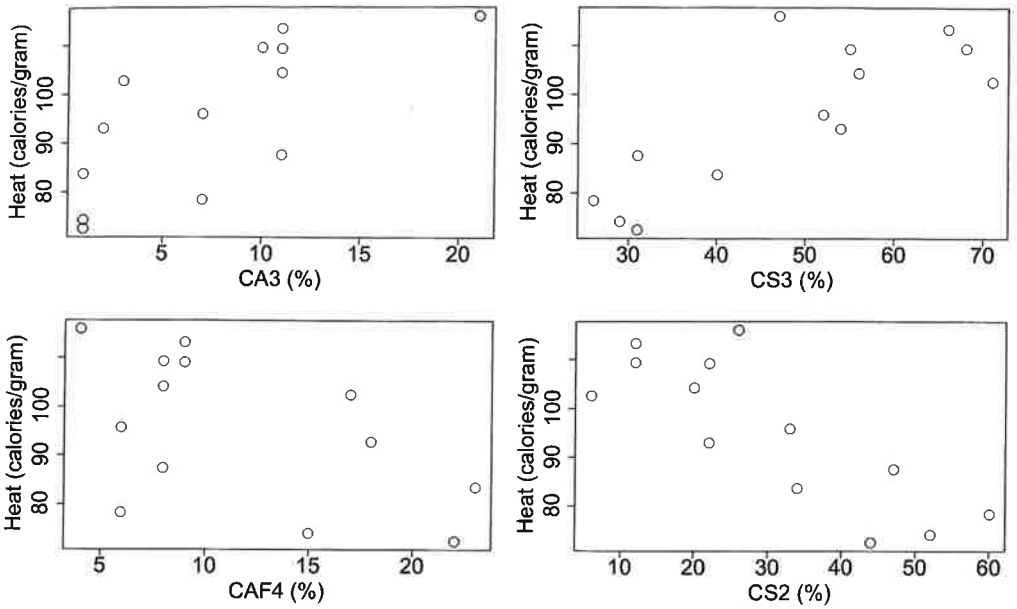
```
> cement <- read.table("cement.txt",header=T)
> attach(cement)
> cement
```

	CA3	CS3	CAF4	CS2	heat
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

The response variable `heat` is the heat evolved in calories per gram of cement, and the explanatory variables are

`CA3`: the percentage weight in clinkers of $3CaO \cdot Al_2O_3$
`CS3`: the percentage weight in clinkers of $3CaO \cdot SiO_2$
`CAF4`: the percentage weight in clinkers of $4CaO \cdot Al_2O_3 \cdot Fe_2O_3$
`CS2`: the percentage weight in clinkers of $2CaO \cdot SiO_2$

```
> plot(CA3,heat)
> plot(CS3,heat)
> plot(CAF4,heat)
> plot(CS2,heat)
```




```
> a <- lm(heat~CA3+CS3+CAF4+CS2)
> summary(a)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
CA3           1.5511     0.7448   2.083   0.0708 .
CS3           0.5102     0.7238   0.705   0.5009
CAF4          0.1019     0.7547   0.135   0.8959
CS2          -0.1441     0.7091  -0.203   0.8441

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-squared:  0.9824,    Adjusted R-squared:  0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

Note that the parameter estimate for each of the explanatory variables in the multiple regression is different from when only regressing `heat` on one of the explanatory variables. For example, if we only use `CA3`, we have

```
> temp <- lm(heat~CA3)
> summary(temp)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  81.4793     4.9273   16.54 4.07e-09 ***
CA3           1.8687     0.5264    3.55  0.00455 **
```

The parameter estimate $\hat{\beta}_1 = 1.5511$ in the multiple regression suggests that, after allowing for the effects of all the other explanatory variables, a unit increase in CA3 results in an increase of 1.5511 calories per gram in heat. Now, after allowing for the effect of CA3 on heat, is there any effect of the rest of the explanatory variables on heat? We can compare the multiple regression and the simple regression by using

```
> anova(temp,a)
```

```
Analysis of Variance Table

Model 1: heat ~ CA3
Model 2: heat ~ CA3 + CS3 + CAF4 + CS2
  Res.Df     RSS Df Sum of Sq     F    Pr(>F)
1     11 1265.69
2      8   47.86  3    1217.8 67.85 4.956e-06 ***
```

This tells us that there is strong evidence that the other explanatory variables affect heat. However, only CA3 has a p-value less than 0.1 in the presence of all the other explanatory variables. We first drop the variable CAF4 which has the highest p-value, and redo the regression.

```
> b <- lm(heat~CA3+CS3+CS2)
> summary(b)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.6483    14.1424   5.066 0.000675 ***
CA3           1.4519     0.1170  12.410 5.78e-07 ***
CS3           0.4161     0.1856   2.242 0.051687 .
CS2          -0.2365     0.1733  -1.365 0.205395
```

The next step, we drop the variable CS2 and redo the regression.

```
> c <- lm(heat~CA3+CS3)
> summary(c)
```

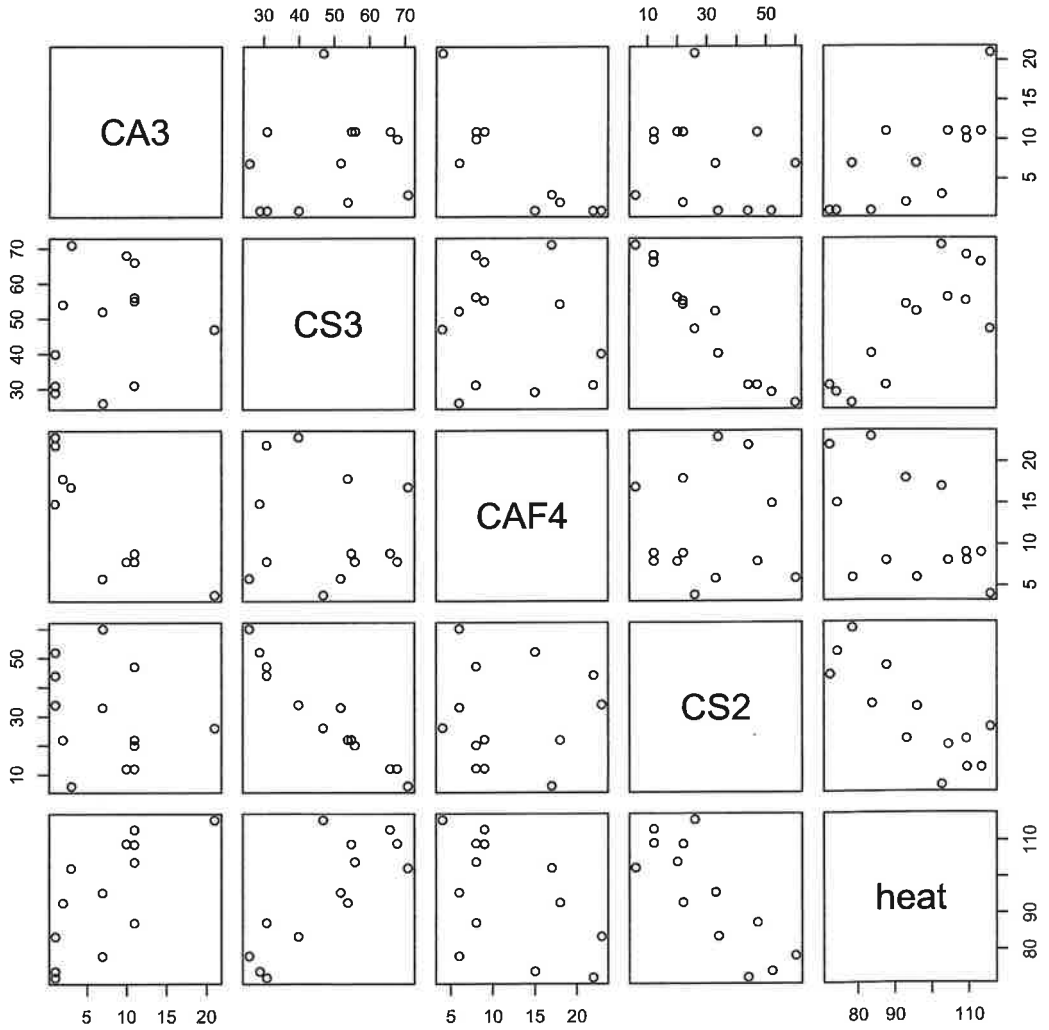
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.57735    2.28617   23.00 5.46e-10 ***
CA3          1.46831    0.12130   12.11 2.69e-07 ***
CS3          0.66225    0.04585   14.44 5.03e-08 ***

Residual standard error: 2.406 on 10 degrees of freedom
Multiple R-squared:  0.9787,    Adjusted R-squared:  0.9744
F-statistic: 229.5 on 2 and 10 DF,  p-value: 4.407e-09
```

By comparing the two models `a` and `c` using ANOVA, we can decide to use Model `c` as our best model.

Collinearity

In fact, there were problems of *collinearity* in the above example – high correlations among the explanatory variables in regression models. Presence of collinearity may result in contradictory conclusions for different samples from the same population. A good way to examine this at the start of our data analysis is to use the function `pairs()` in R to plot the data.

```
> pairs(cement)
```



When an explanatory variable is correlated with a combination of other explanatory variables, a good way to identify collinearity is to calculate the variance inflation

factors (VIFs)

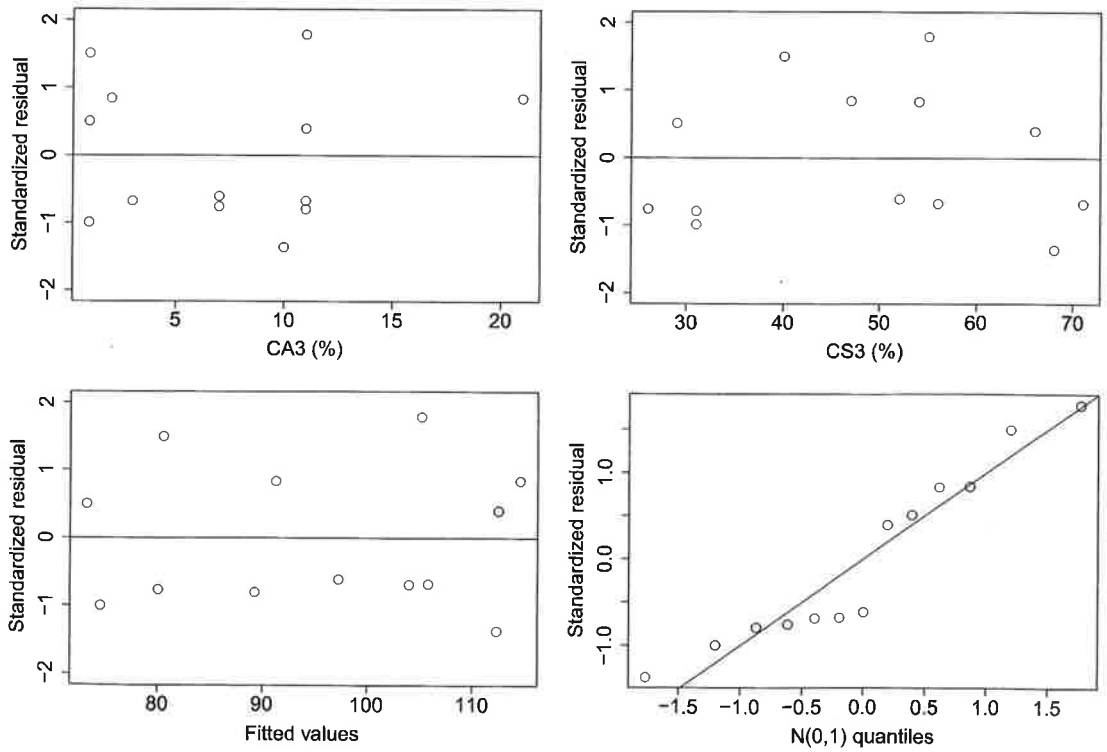
$$VIF(i) = \frac{1}{1 - R^2(i)}$$

where $R^2(i)$ is the R^2 value when we regress the explanatory variable x_i on all the other explanatory variables. Normally, if the VIFs are greater than 10, then there is a collinearity problem. In R, we can use the function `vif()` to calculate the VIFs.

```
> library(car)
> vif(a)
```

CA3	CS3	CAF4	CS2
38.49621	254.42317	46.86839	282.51286

```
> r <- rstandard(c)
> lf <- fitted(c)
> plot(CA3,r,xlab="CA3 (%)",ylab="Standardized residual")
> abline(h=0)
> plot(CS3,r,xlab="CS3 (%)",ylab="Standardized residual")
> abline(h=0)
> plot(lf,r,xlab="Fitted values",ylab="Standardized residual")
> abline(h=0)
> qqnorm(r,xlab="...",ylab="...",main="")
> abline(0,1)
```



Categorical explanatory variables

Sometimes we will encounter situations in which some of the explanatory variables are categorical, and others are continuous.

We look at the example of comparing the regrowth of Ipomopsis aggregata following removal of its primary shoot by herbivores (Crawley, 2012, *The R Book*, John Wiley & Sons, Ltd.). The grazed group has twenty plants. They were exposed to rabbits during the first two weeks of stem elongation, and then protected from subsequent grazing and allowed to regrow. Another 20 plants were not grazed. Before potting each plant, the diameter of the top of the rootstock (in millimeter, `root`) was measured. We are interested in how the variable `root` influence the fruit production (dry weight in milligrams, `fruit`).

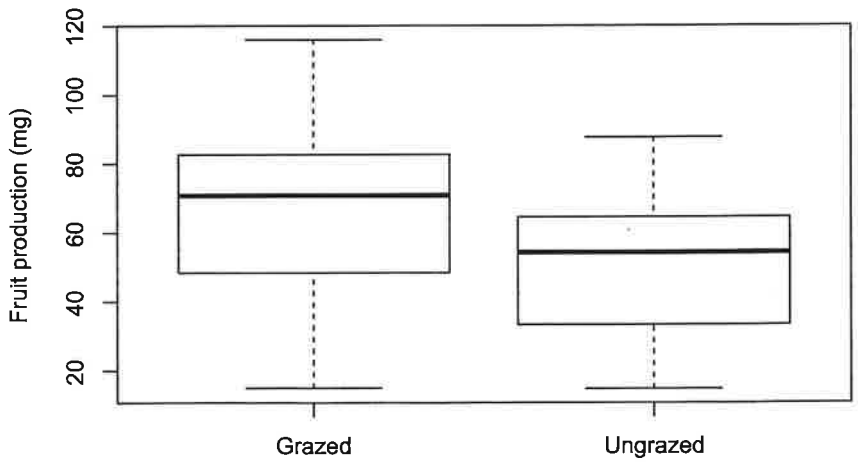
```
> plant <- read.table("grazingplant.txt",header=T)
> attach(plant)
> names(plant)
```

```
[1] "root"    "fruit"   "grazing"
```

Common mistakes:

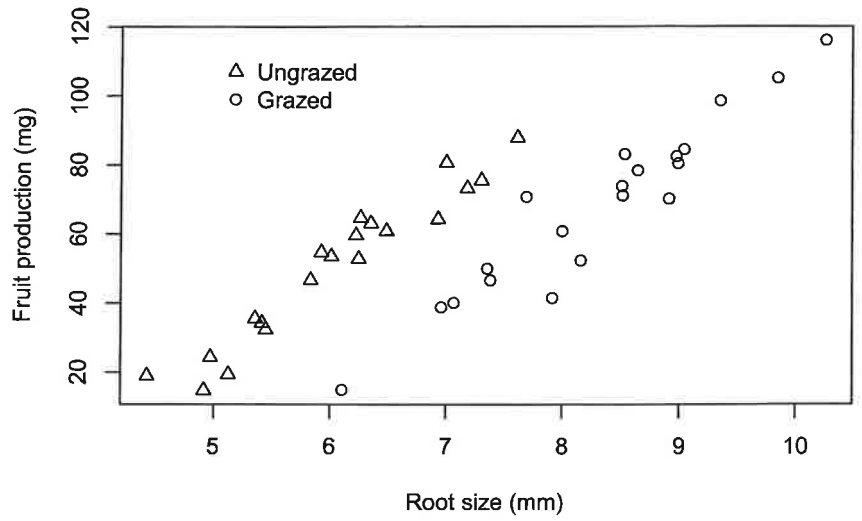
```
> grazing <- factor(grazing)
> boxplot(fruit~grazing)
```

define variable as a factor.



By examining this plot, it seems that the fruit production is higher for grazed plants. Is this really the case? Notice that there is another variable `root` which may affect the fruit production as well.

```
> plot(root,fruit,pch=as.numeric(grazing))
> legend(5,115,unique(grazing),
        pch=unique(as.numeric(grazing)),bty="n")
```



This scattered plot shows that the fruit production for the ungrazed plants is higher than that for the grazed plants with the same initial root size. To analyze the effects of both grazing and root size on the fruit production, we use the following model.

```
> a <- lm(fruit~root+grazing)
> summary(a)
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -127.829      9.664  -13.23 1.35e-15 ***
root              23.560      1.149   20.51  < 2e-16 ***
grazingUngrazed   36.103      3.357   10.75 6.11e-13 ***

Residual standard error: 6.747 on 37 degrees of freedom
Multiple R-squared:  0.9291,    Adjusted R-squared:  0.9252
F-statistic: 242.3 on 2 and 37 DF,  p-value: < 2.2e-16
```

Here, `grazing` is treated as a dummy indicator variable. If we define

$$g = \begin{cases} 1, & \text{ungrazed group} \\ 0, & \text{grazed group} \end{cases}$$

and let x_i be the ith observation of `root`, then fitting the model $y_i = \beta_0 + \beta_1 x_i + \beta_2 g_i + \varepsilon_i$, we will get $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 g$. This is equivalent to

$$\begin{cases} \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2, & \text{ungrazed group} \\ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, & \text{grazed group} \end{cases}$$

From the table of parameter estimates, we can conclude that for a given root size, the ungrazed plants produce 36.103mg of fruits more than the grazed plants, with a 95% confidence interval (29.301mg,42.906mg). Within the group of grazed (or ungrazed) plants, increasing the root size by 1mm will increase the fruit production by 23.560mg with a 95% confidence interval (21.232mg,25.888mg).

```
> confint(a)
```

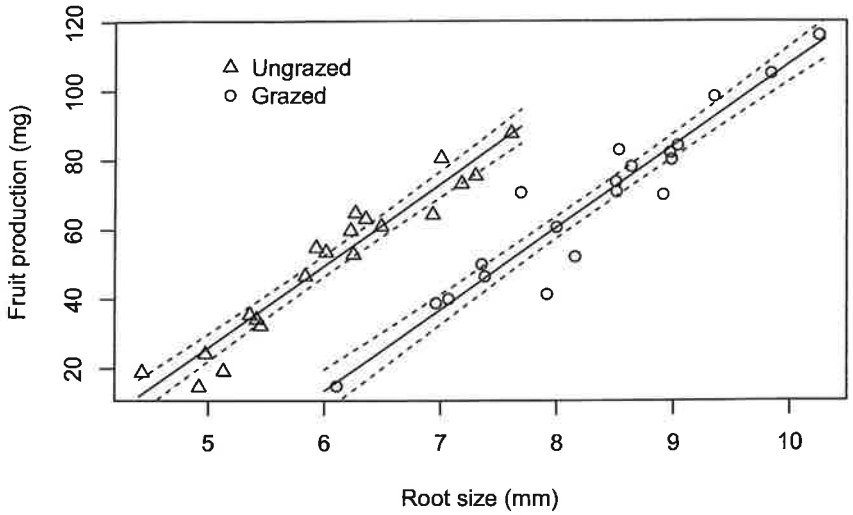
```
                    2.5 %     97.5 %
(Intercept)     -147.41068 -108.24804
root              21.23242   25.88768
grazingUngrazed   29.30052   42.90598
```

```
> anova(a)
```

```
Analysis of Variance Table

Response: fruit
          Df  Sum Sq Mean Sq F value    Pr(>F)
root       1 16795.0 16795.0  368.91 < 2.2e-16 ***
grazing    1  5264.4  5264.4  115.63 6.107e-13 ***
Residuals 37  1684.5    45.5
```

We use model a to predict the fitted vales for a smooth range of `root` values for the grazed and ungrazed groups with the confidence intervals of the predicted values.

```
> plot(root,fruit,pch=as.numeric(grazing))
> legend(5,115,unique(grazing),
        pch=unique(as.numeric(grazing)),bty="n")
> x <- seq(from=6,to=10.3,by=0.1)
> g <- rep("Grazed",length(x))
> x1 <- seq(from=4.4,to=7.7,by=0.1)
> g1 <- rep("Ungrazed",length(x1))
> y <- predict(a,list(root=x,grazing=g),interval="confidence")
> y1 <- predict(a,list(root=x1,grazing=g1),interval="confidence")
> matlines(x,y,lty=c(1,2,2),col="black")
> matlines(x1,y1,lty=c(1,2,2),col="black")
```




```
> r <- rstandard(a)
> lf <- fitted(a)
> plot(lf,r,xlab="Fitted values",ylab="Standardized residual")
> abline(h=0)
> qqnorm(r,xlab="N(0,1) quantiles",ylab="Standardized residual")
> abline(0,1)
```