**Linear Models**

We have seen examples where there are two variables, x and y, that are related (there is structure to the scatter plot), but the relationship is not deterministic; i.e., the value of x does not determine the value of y. Often the relation between x and y is partly deterministic and partly “random”. For example, there is a connection between one’s score on the ACT math test and one’s grade in Calculus 1, but it is not deterministic. Your ACT math score doesn’t determine your Calculus 1 grade. A model for the relation between x = ACT math score and y = grade (on a 4 pt. scale) in Calculus 1 might look like this:

y = f(x) + 

where f(x) is an ordinary (deterministic) function and  is a “random” contribution to an individual’s Calculus 1 grade. You might think of  as a summary of other factors that affect a Calculus 1 grade that are independent of ACT math score.

**Another Example**: If you are making repeated measurements of the mass of an object, you are likely to get slightly different measurement values for the same object. If x is the true value of the mass, and y is the value of the measurement, then y = f(x) + , where f(x) = x is the true mass and  represent the measurement error, which will be different for different measurements of the object.

**Another2 Example**: The admissions office to Alvin University requires applicants to submit SAT test scores and high school GPA. The university thinks (with justification) that these numbers are useful in predicting a student’s university GPA at the end of the freshman year. There are 3 variables involved here: x1 = SAT score, x2 = high school GPA, and y = university freshman GPA. The variables x1 and x2 are **explanatory variables** and y is the **response variable**. The admissions office knows that the relationship involved is not of the form y = f(x1,x2) (deterministic). Rather, it considers the relationship to have the form

Y = f(x1,x2) + 

where  varies from one student to another, but is independent of the student’s SAT score and high school GPA.

**Simple Linear Regression Model**

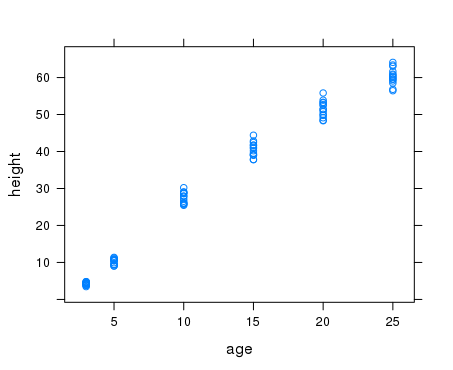
A simple linear regression model with one explanatory variable x has the form

Y= , where  is a random variable that is N(0, ).

**Note that Y is a random variable, not a single number.**

**Example**: The data-frame Loblolly contains data on the height and age of 84 loblolly pine trees. In general, how does the age of a loblolly pine affect its height? Here, y = height is the response variable and x = age is the explanatory variable.

* xyplot(height~age,data=Loblolly)



**Remember:** In the R command we use the response variable as the first column and the explanatory variable as the second. That puts the explanatory variable on the horizontal axis and the response variable on the vertical axis.

You can “see” in this graph that the height seems to depend linearly on the age, but that there is also a random  component as well. From the graph, it looks like a simple linear regression model might fit well; i.e.,

Height = 

for some values of , where is N(0, ).

**NOTE** that the intent is that the model describe the relation between age and height for **ALL** loblolly pines, not just the trees in the sample.

There are three parameters in the model

is the **intercept parameter**

 is the **slope parameter**

 is the **standard deviation parameter.**

How can we use the data to **estimate** the values of these parameters?

**Estimating the linear portion of the model**: Use the line that “best fits” the data in the least squares sense.

**Example** Given the three points (1,1), (2,3), and (4,5), what line best fits these points in the least squares sense?

**We won’t do that!**

Below are the general (**UGLY**) formulas for the coefficients of the least squares line for a general set of data points.

**General formulas**: If the data is (x1,y1), … , (xn,yn) then the coefficients of the least squares line

y = b0 + b1x are given by

b0 =

b1 = 

Happily, R does the computations for us.

> lm(height~age,data=Loblolly)

Call:

lm(formula = height ~ age, data = Loblolly)

Coefficients:

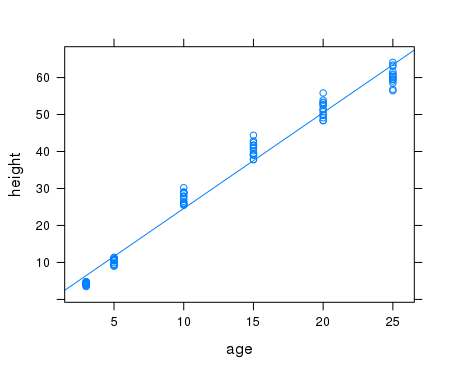
(Intercept) age

-1.312 2.591

**So, the least squares line is y**

We can add the regression line to the scatterplot

> xyplot(height~age,data=Loblolly,type=c("p","r"))



**Correlation Coefficient** r (x,y) = 

**The correlation coefficient measures the strength of linear association between the variables x and y.**

1. |r| ≤ 1
2. r > 0 corresponds to positive linear association
3. r < 0 corresponds to negative linear association
4. Small |r| corresponds to little if any linear association
5. Large |r| corresponds to large linear association
6. A linear change in the units of measurement in x and/or y does not change the value of r.
7. r(x,y) = r(y,x)

**Calculating the correlation coefficient using R**

cor(col1,col2,data=dataframe)

**Alternate formulas for the regression coefficients**

**Example**

> sh<-sd(~height, data=Loblolly)

> sa<-sd(~age, data=Loblolly)

> r<-cor(height,age,data=Loblolly)

> sh

[1] 20.6736

> sa

[1] 7.899977

> r

[1] 0.9899132

> r\*sh/sa

[1] 2.590523

**Estimating the “error” term  and checking for normality**

**Residuals**

If (xi ,yi ) is one of the data points and y = b0 + b1 x is the regression line, then the difference between the observed y : yi and the predicted y:  = b0 + b1 xi is the ith **residual**. We can think of the residuals as values of the  term in the model. If the model is appropriate, then the residuals should have a normal distribution.

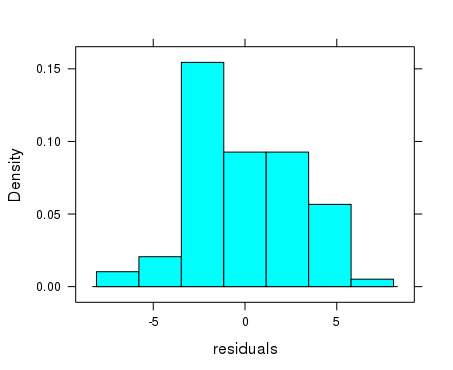
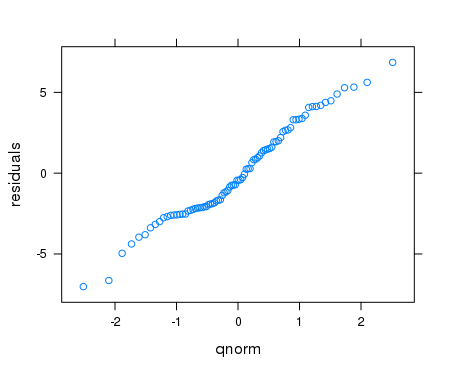
**residual = observed – predicted.**

Example:

> residuals<-resid(lm(height~age,data=Loblolly))

> histogram(residuals)

> qqmath(residuals)



We can also use the residuals to estimate the standard deviation  of the error term. The formula for the estimation of  is , where the ei’s are the residuals.

**Again, R comes to the rescue for the computations.**

> summary(lm(height~age,data=Loblolly))

Call:

lm(formula = height ~ age, data = Loblolly)

Residuals:

Min 1Q Median 3Q Max

-7.0207 -2.1672 -0.4391 2.0539 6.8545

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.31240 0.62183 -2.111 0.0379 \*

age 2.59052 0.04094 63.272 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Residual standard error: 2.947** on 82 degrees of freedom

Multiple R-squared: 0.9799**,** Adjusted R-squared: 0.9797

F-statistic: 4003 on 1 and 82 DF, p-value: < 2.2e-16

**The Residual standard error is the estimate of , the standard deviation of the error term.**

**Why is R2 given instead of R?**

**For example: If the correlation coefficient between height (explanatory) and weight (response) were r = .7, then r2 = .49 would mean that**

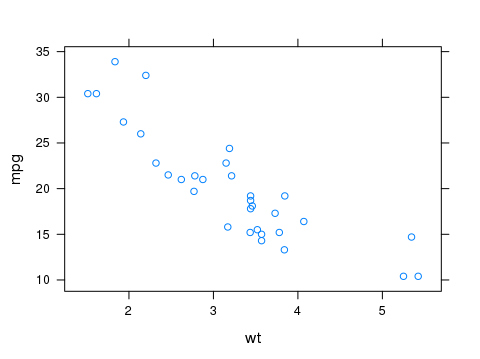
**Example** The data-frame **mtcars** contains **mpg** and **wt** data on 32 automobile models from 1974. We would expect the weight of a vehicle to affect its fuel efficiency. Use this data to explore the relation between weight and fuel efficiency.

**Explanatory variable:**

**Response variable**

**Scatter plot**

> xyplot(mpg~wt,data=mtcars



**The simple linear regression model (estimated)**

* summary(lm(mpg~wt,data=mtcars))

Call:

lm(formula = mpg ~ wt, data = mtcars)

Residuals:

Min 1Q Median 3Q Max

-4.5432 -2.3647 -0.1252 1.4096 6.8727

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) **37.2851** 1.8776 19.858 < 2e-16 \*\*\*

wt **-5.3445** 0.5591 -9.559 1.29e-10 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: **3.046** on 30 degrees of freedom

Multiple R-squared: 0.7528, Adjusted R-squared: 0.7446

F-statistic: 91.38 on 1 and 30 DF, p-value: 1.294e-10

**MPG = +  where  is N(0, ?)**

**Also, since r2 = .7528,**

**Check for normality**

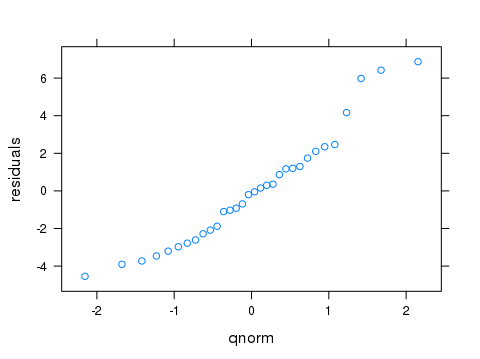
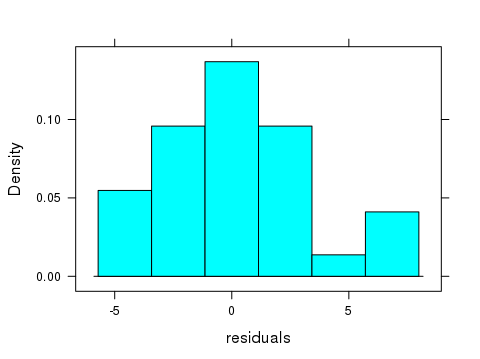
**We can get the residuals for a linear model using:**

**residuals<- resid(lm(y~x,data = dataframe))**

> residuals<-resid(lm(mpg~wt, data = mtcars))

> histogram(residuals)

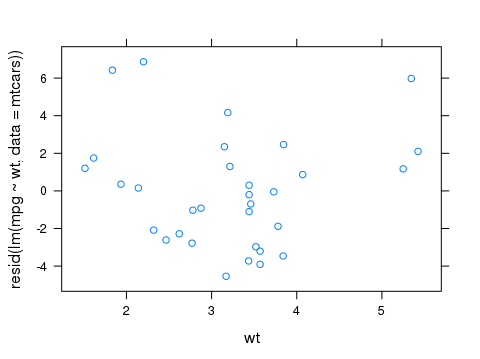
> qqmath(residuals)



**One other issue to check**: In addition to the assumption that the error term  is normal, the model assumes that the variability of the response is independent of x. In terms of the data, the variability of the residuals should be independent of x. One way to check this is to plot the residuals with respect to x. If the residuals have the same spread with respect to 0 for all x’s that is good. If the spread with respect to 0 varies a lot with respect to x, that is bad.

**Using R**

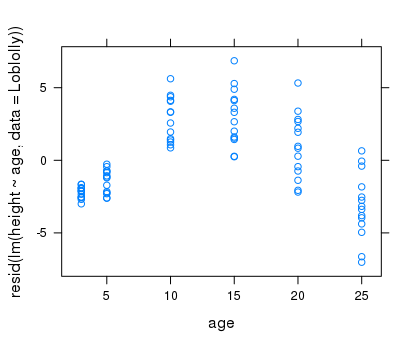
* xyplot(resid(lm(mpg~wt,data=mtcars))~wt,data=mtcars)



**“Good” plot?**

**Check the residual plot for loblolly pines**

* xyplot(resid(lm(height~age,data=Loblolly))~age,data=Loblolly)

****

**“Good” plot?**

**Exercises 14**

1. The data-frame **cars** contains 50 measurements of speed (**speed**) and stopping distance (**dist**). Note that this data is from the 1920’s.
2. Which of the two variables should be considered the explanatory variable and which should be considered the response variable?
3. Create a scatterplot for this data. Does it look like there is linear association between these variables?
4. Find the least squares regression line for this data. Include the R commands and output as well as the equation of the regression line.
5. Recreate the scatterplot so that it includes the regression line.
6. The data-frame **trees** contains the **Height**, **Volume**, and **Girth** for 31 cherry trees.
7. Let Girth be the explanatory variable and Volume the response variable. Compute the least squares regression line: Volume = b0 + b1\*Girth.
8. Solve the equation in (a) for Girth in terms of Volume.
9. Let Volume be the explanatory variable and Girth the response variable. Compute the least squares regression line: Girth = a0 + a1Volume.
10. Are the two equations in (b) and (c) the same?
11. Use the R command

square<-data.frame(x=c(-3,-2,-1,0,1,2,3),y=c(9,4,1,0,1,4,9))

to create the data-frame **square:**

|  |
| --- |
| x y  1 -3 9  2 -2 4  3 -1 1  4 0 0  5 1 1  6 2 4  7 3 9 |
|  |

1. Produce the scatter plot for this data. Does it look like there is linear association between x and y? Is there non-linear association?
2. Find the least squares regression line and reproduce the scatter plot with the regression line.
3. Find the correlation coefficient for this data.
4. Comment on the truth of the statement : If the correlation coefficient is 0, there is no association/connection between x and y.
5. The data-frame **cars** contains 50 measurements of speed (**speed**) and stopping distance (**dist**).
6. Produce a histogram of the residuals and a normal qq plot of the residuals. Does it look like the assumption that the error term has a normal distribution is (at least somewhat) plausible?
7. What is the estimate of the standard deviation of the error term?
8. The data-frame **faithful** contains data on **eruptions** and **waiting.**  One might expect that there is an association between the length of an eruption and the waiting time until the next eruption. Let eruptions be the explanatory variable and waiting the response variable.
9. Create a scatter plot for this data that includes a plot of the regression line. Does it look like a simple linear regression model would be a good model? Is the association positive or negative? Include the scatter plot with your answers.
10. Create a histogram of the residuals and a qqmath plot of the residuals. Does it look like the normality assumption for a SLR model holds? Include the plots with your answer.
11. Create a residual plot of the residuals vs eruptions. Does it look like the equal variance assumption holds? Include the plot with your answer.
12. Find the correlation coefficient.
13. Find the estimated SLR model from this data. Write it in the form

**waiting = ? + ?\*eruptions +** , where  is N(0,?)

1. What percent of the variability in waiting time is explained by the differences in eruption times?