Chap 4

## Simple linear models

The linear model is given by

* are random (need some assumptions)
* are **fixed** (*independent/preditor* variable)
* are random (*dependent/response* variable)
* is the *intercept*
* is the *slope*

### Least square estimators

Choose to minimize

The minimizers satisfy

This gives

Define

We thus have

Regression function: .

### Expected values

Assumption A1: .

Theorem 1: Under Assumption A1, are unbiased estimators for , respectively.

Proof:

### Variances

Assumption A2: .

Theorem 2: Under Assumption A2, we have

Proof: Since for any , . We thus have

We next show that .

So bigger is better. Get a bigger sample size if you can. Smaller is better. The most interesting one is that bigger is better. The more spread out the are the better we can estimate the slope . When you’re picking the , if you can spread them out more, then it is more informative.

### Estimation of

For Assumption A2, it is common that the variance is unknown. The next theorem gives an unbiased estimate of .

Definition: The residual sum of squares (RSS) is defined by

Theorem 3: Let

Under Assumptions A1 and A2, we have .

Proof: Let .

As a result, we have

### Normal distributions

Assumption B: .

Assumption B includes Assumptions A1 and A2.

Theorem 4: Under Assumption B, we have

(1).

(2).

(3).

(4). is independent of .

Proof: Under Assumption B, independently. Both are linear combinations of s. Consequently, they are normally distributed. We have known their expected values and variances from Theorems 1 and 2. The claims (1) and (2) are thus verified. The proofs of claims (3) and (4) are deferred to the general case.

It is degrees of freedom because we have fit two parameters to the data points.

### Confidence intervals and hypothesis tests

For known we can make tests and confidence intervals using

The confidence interval for is given by . For testing

we reject if with the most popular hypothesized value being (i.e., the regession function is **significant** or not at significance level .)

In the more realistic setting of unknown , so long as , using claims (2-4) gives

The confidence interval for is . For testing

we reject if .

For drawing inferences about , we can use

The confidence interval for is

### Case study 1

A manufacturer of air conditioning units is having assembly problems due to the failure of a connecting rod to meet finished-weight specifications. Too many rods are being completely tooled, then rejected as overweight. To reduce that cost, the company’s quality-control department wants to quantify the relationship between the weight of the **finished rod**, , and that of the **rough casting**, . Castings likely to produce rods that are too heavy can then be discarded before undergoing the final (and costly) tooling process. The data are displayed below.

rough weight vs. finished weight

id

rough\_weight

finished\_weight

1

2.745

2.080

2

2.700

2.045

3

2.690

2.050

4

2.680

2.005

5

2.675

2.035

6

2.670

2.035

7

2.665

2.020

8

2.660

2.005

9

2.655

2.010

10

2.655

2.000

11

2.650

2.000

12

2.650

2.005

13

2.645

2.015

14

2.635

1.990

15

2.630

1.990

16

2.625

1.995

17

2.625

1.985

18

2.620

1.970

19

2.615

1.985

20

2.615

1.990

21

2.615

1.995

22

2.610

1.990

23

2.590

1.975

24

2.590

1.995

25

2.565

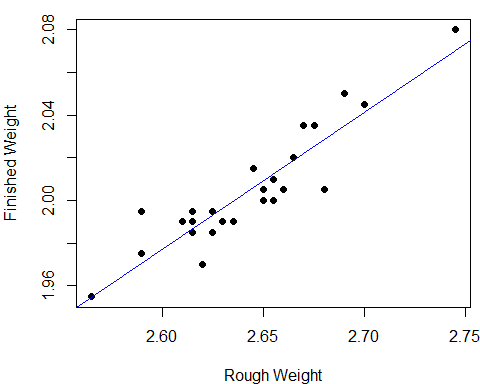
1.955

Consider the linear model

The observed data gives , , , , . The least square estimates are

The regession function ; see the blue line given below.

attach(rod)  
par(mar=c(4,4,1,0.5))  
plot(rough\_weight,finished\_weight,type="p",pch=16,  
 xlab = "Rough Weight",ylab = "Finished Weight")  
lm.rod = lm(finished\_weight~rough\_weight)  
abline(coef(lm.rod),col="blue")



summary(lm.rod) #output the results

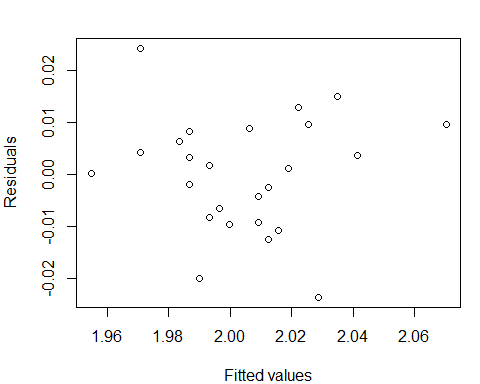
##   
## Call:  
## lm(formula = finished\_weight ~ rough\_weight)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.023558 -0.008242 0.001074 0.008179 0.024231   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.30773 0.15608 1.972 0.0608 .   
## rough\_weight 0.64210 0.05905 10.874 1.54e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01131 on 23 degrees of freedom  
## Multiple R-squared: 0.8372, Adjusted R-squared: 0.8301   
## F-statistic: 118.3 on 1 and 23 DF, p-value: 1.536e-10

### Assessing the Fit

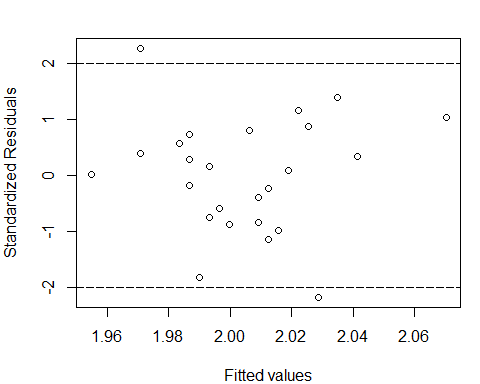
As an aid in assessing the quality of the fit, we will make extensive use of the residuals, which are the differences between the observed and fitted values:

It is most useful to examine the residuals graphically. Plots of the residuals versus the values may reveal systematic misfit or ways in which the data do not conform to the fitted model. Ideally, the residuals should show no relation to the values, and the plot should look like a horizontal blur. The residuals for case study 1 are plotted below.

par(mar=c(4,4,2,1))  
plot(lm.rod$fitted.values,lm.rod$residuals,"p",  
 xlab="Fitted values",ylab = "Residuals")

 Standardized Residuals are graphed below. The key command is rstandard.

par(mar=c(4,4,2,1))  
plot(lm.rod$fitted.values,rstandard(lm.rod),"p",  
 xlab="Fitted values",ylab = "Standardized Residuals")  
abline(h=c(-2,2),lty=c(5,5))



### Drawing Inferences about

For given , we want the estimate the expected value of , i.e., A natural unbiased estimate is . From the proof of Theorem 3, we have the variance

Under Assumption B, by Theorem 4, we have

We thus have the following results.

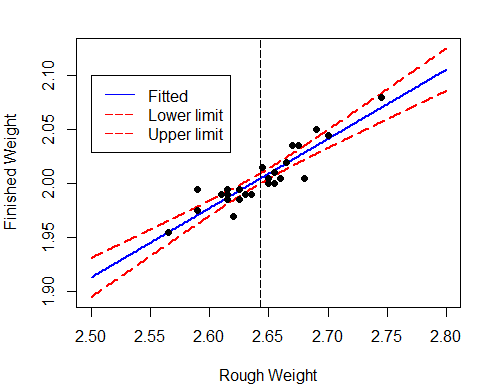
Theorem 5: Suppose Assumption B is satisfied. Then we have

A confidence interval for is given by

Notice from the formula in Theorem 5 that the width of a confidence interval for increases as the value of becomes more extreme. That is, we are better able to predict the location of the regression line for an -value close to than we are for -values that are either very small or very large.

For case study 1, we plot the lower and upper limits for the confidence interval for .

x = seq(2.5,2.8,by=0.001)  
newdata = data.frame(rough\_weight= x)  
pred\_x = predict(lm.rod,newdata,interval = "confidence")  
par(mar=c(4,4,2,1))  
matplot(x,pred\_x,type="l",lty = c(1,5,5),  
 col=c("blue","red","red"),lwd=2,  
 xlab="Rough Weight",ylab="Finished Weight")  
abline(v=mean(rough\_weight),lty=5)  
points(rough\_weight,finished\_weight,pch=16)  
legend(2.5,2.1,c("Fitted","Lower limit","Upper limit"),  
 lty = c(1,5,5),col=c("blue","red","red"))



### Drawing Inferences about Future Observations

We now give a **prediction interval** for the future observation rather than its expected value . Note that here is no longer a fixed parameter, which is assumed to be independent of ’s. A prediction interval is a range of numbers that contains with a specified probability. Consider . If Assumption A1 is satisfied, then

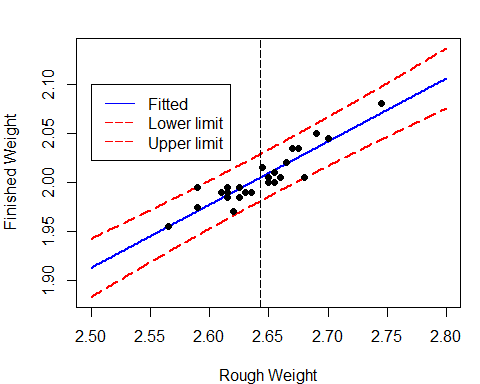
If Assumption A2 is satisfied, then

If Assumption B is satisfied, is then normally distributed.

Theorem 6: Suppose Assumption B is satisfied. Let , where is independent of ’s. A prediction interval for is given by

For case study 1, we plot the lower and upper limits for the prediction interval for .

x = seq(2.5,2.8,by=0.001)  
newdata = data.frame(rough\_weight= x)  
pred\_x = predict(lm.rod,newdata,interval = "prediction")  
par(mar=c(4,4,2,1))  
matplot(x,pred\_x,type="l",lty = c(1,5,5),  
 col=c("blue","red","red"),lwd=2,  
 xlab="Rough Weight",ylab="Finished Weight")  
abline(v=mean(rough\_weight),lty=5)  
points(rough\_weight,finished\_weight,pch=16)  
legend(2.5,2.1,c("Fitted","Lower limit","Upper limit"),  
 lty = c(1,5,5),col=c("blue","red","red"))

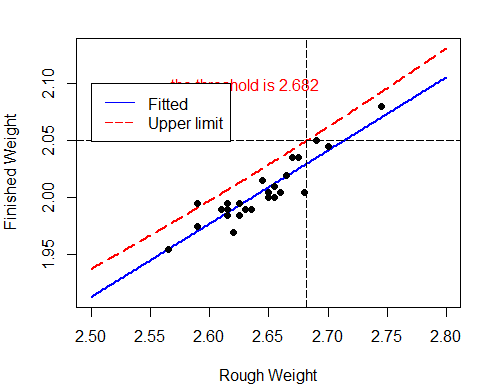


### How to control y?

Consider case study 1 again. Castings likely to produce rods that are too heavy can then be discarded before undergoing the final (and costly) tooling process. The company’s quality-control department wants to produce the rod with weights no large than 2.05. How to choose the rough casting?

Now we want with probability . Similarly to Theorem 6, we can construct one-side confidence interval for , that is

This implies



## Multiple linear regression

With problems more complex than fitting a straight line, it is very useful to approach the linear least squares analysis via linear algebra. Consider a model of the form

Let , , , and let be the matrix

The model can be rewritten as

* the matrix is called the **design matrix**,
* assume that .

The least squares problem can then be phrased as follows: Find to minimize

where is the Euclidean norm.

Note that

If we differentiate with respect to each and set the derivatives equal to zero, we see that the minimizers satisfy

We thus arrive at

If the design matrix is **nonsingular**, the formal solution is

The following lemma gives a criterion for the existence and uniqueness of solutions of the normal equations.

Lemma 1: The design matrix is nonsingular if and only if .

Proof: First suppose that is singular. There exists a nonzero vector such that . Multiplying the left-hand side of this equation by , we have So , the columns of are linearly dependent, and the rank of is less than .

Next, suppose that the rank of is less than so that there exists a nonzero vector such that . Then , and hence is singular.

In what follows, we assume that .

### Expected values and variances

Assumption A: Assume that and .

Theorem 7: Suppose that Assumption A is satisfied and , we have

(1).

(2). ,

Proof:

We used the fact that for any fixed matrix , and and therefore are symmetric.

### Estimation of

The residual sum of squares (RSS) is defined by

where the vector of residuals is

* is an matrix (called the **projection matrix**).

Two useful properties of are given in the following lemma.

Lemma 2: Let be defined as before. Then

We may think geometrically of the fitted values, $Y=X$, as being the projection of onto the subspace spanned by the columns of .

The sum of squared residuals is then

Lemma 3: The cyclic property of the trace, that is, .

Using Lemma 3, we have

where we used . Using the cyclic property of the trace again gives

We therefore have .

Theorem 8: Suppose that Assumption A is satisfied and ,

is an unbiased estimate of .