**Interpreting and Using Linear Regression Models**

1. The estimated model is generally effective only for values of the explanatory variable that are fairly close to the values of x that occurred in the data used to estimate the model. **Extrapolating** outside the range of the data used to estimate the model is dangerous.

2. The slope parameter, in general, is important. It tells us the rate at which the response is changing with respect to the explanatory variable.

**Example**: The regression equation for the data mtcars is

mpg = 37.2851 + -5.3445 wt.

The slope = -5.3445 tells us that, on average, an increase in the weight of a car by 1000 pounds reduces its fuel efficiency by more than 5 miles per gallon.

If the slope parameter were 0, that would say that there is no linear association between the explanatory variable and the response. In this case, Y = + , and the response is independent of x. So, whether the slope coefficient is 0 or not is important information about the linear relation between the two variables. Of course, we don’t know the actual value of the slope, we only have the estimated value based on our data. Our estimate might be non-zero when the actual value is zero. So, we could set up a hypothesis test about the slope,

H0 :  = 0

Ha :  ≠ 0

and calculate the p-value of the data for these hypotheses. Happily, R does this as part of the summary command.

> 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

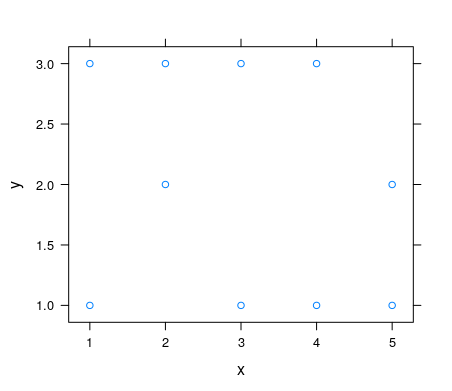
Multiple R-squared: 0.7528, Adjusted R-squared: 0.7446

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

**Example: Here is an artificial example**

> Example<-data.frame(x=c(1,1,2,2,3,3,4,4,5,5),y=c(1,3,2,3,1,3,1,3,1,2))

> xyplot(y~x,data = Example)



We can see from the scatterplot, that the y (response) value doesn’t seem to depend on x (explanatory). So, we would expect that the slope of the linear model, We let R produce the estimated linear model, , would be 0. Is there strong evidence that the slope is **not** 0? We use R to find the estimated linear model.

> summary(lm(y~x,data=Example))

Call:

lm(formula = y ~ x, data = Example)

Residuals:

Min 1Q Median 3Q Max

-1.3000 -0.8125 0.0750 0.8125 1.1500

Coefficients:

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

(Intercept) 2.4500 0.7205 3.401 0.00935 \*\*

x **-0.1500** 0.2172 -0.691 0.50940

---

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

Residual standard error: 0.9715 on 8 degrees of freedom

Multiple R-squared: 0.05625, Adjusted R-squared: -0.06172

F-statistic: 0.4768 on 1 and 8 DF, p-value: 0.5094

We see that the estimate of the slope (-0.15) is close to 0 and the p-value is about 51%. **So, the data does not provide strong evidence that the slope is non-zero.**

3. The intercept parameter is often not interesting in itself. It gives the response when the explanatory variable = 0. The value x = 0 may not be close to the data values used or x = 0 may not even be a sensible value for x.

**Example:** The intercept value 37.2851 in the mtcars example would be

**Example**: The regression equation for height~age (where height is measured in feet and age is measured in years) for loblolly pines is

height = -1.31240 + 2.59052 age

What is the significance of the slope parameter?

What is the significance of the intercept parameter?

The data involved trees whose height was measured at 3,5,10,15,20,25 years. Which of the following ages (x) would you expect this model to be useful for: 8, 17,23,28,45?

4. **Confidence intervals for the ’s**

R will produce CI’s for the slope and intercept parameters

**confint(lm(y~x,data = dataframe), level = )**

> confint(lm(mpg~wt,data=mtcars), level = .9)

5 % 95 %

(Intercept) 34.098303 40.471950

wt -6.293412 -4.395531

**Using a linear regression model**

Let Y = 2 + 3x + , where  is N(0,4), be a simple linear regression model. If x = 5, the response is a random variable Y = 17 + . What is the distribution of Y? Since  has a normal distribution Y = 17 +  has a normal distribution.

E(Y) = **E(17 + ) = 17 + E() = 17 + 0 = 17**

Var (Y) = **Var(17 + ) = Var() = 42 = 16**

So, Y is **N( 17 , 4 ).**

**In general**, if , where  is N(0, ), then for any particular x, the response Y is N(,). The line y =  is called the **line of means**, since for any x it gives the average response to x.

**Estimating the response to a value of x**

There are two kinds of estimations we might be interesting in making about the response to a particular value of x.

**Estimation Problem 1** For a given value of x, estimate the average value of all responses to x.

**Estimation Problem 2** For a given value of x, estimate the value of a particular response to x.

**Example**: Let x = **eruptions** = length of an eruption of Old Faithful and y = **waiting** = time until the next eruption. From the data-frame **faithful,** the estimated linear regression model is

waiting = 33.47 + 10.73 eruptions + , where  is N(0, 5.91)

EP1 Estimate the average waiting time after eruptions that lasted 5 minutes.

**.**

EP2 You have just observed an eruption that lasted 5 minutes. Estimate how long you will have to wait until the next eruption starts.

Which of these estimates do you feel more secure about the accuracy of?

**Confidence interval** for an average response using R

> mod<-makeFun(lm(waitingtime~eruptions,data=faithful))

> mod(5,interval="confidence",level=.95)

fit lwr upr

1 87.1226 85.94933 88.29588

**Prediction interval** for an individual response

> mod(5,interval="prediction",level=.95)

fit lwr upr

1 87.1226 75.4202 98.825

**NOTES:** (1) The center of both intervals, the CI and the PI, is 87.12.

(2) The margin of error in the prediction interval for an individual response is much larger than the margin of error for a confidence interval for the average response.

**The formulas**

data (x1,y1),…,(xn,yn)

 = average of x’s

 = average of y’s

s = estimate of 



**For a confidence interval for an average response** to x = x\*

ME = , where tc depends on the confidence level and n

**For a prediction interval for an individual response** to x = x\*

ME = 

**Comparing the Margins of Error**

1. The ME for a confidence interval is always smaller than the ME for a prediction interval.
2. By increasing the sample size, the ME for a CI can be made arbitrarily small, but the ME for a PI cannot be less than tcs.
3. The closer the value x\* is to , the smaller the ME is for both CI’s and PI’s.
4. The larger the value of SSxx the smaller the ME. What makes SSxx large?

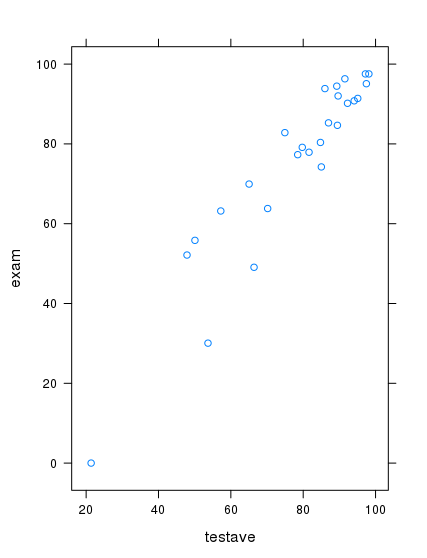
**Example:** It seems reasonable to think that there is an association between a student’s performance on tests and the student’s performance on the final exam. Let’s explore this using a sample of Calculus I students.

Explanatory variable = average % on tests = **testave**

Response variable = % on exam = **exam**

My sample has size n = 26 and is contained in the data-frame class.

* xyplot(exam~testave, data=class)



> summary(lm(exam~testave,data=class))

Call:

lm(formula = exam ~ testave, data = class)

Residuals:

Min 1Q Median 3Q Max

-17.753 -3.407 -1.119 5.550 12.156

Coefficients:

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

(Intercept) -13.90148 6.71159 -2.071 0.0493 \*

testave 1.14968 0.08386 13.709 7.58e-13 \*\*\*

---

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

Residual standard error: 7.969 on 24 degrees of freedom

Multiple R-squared: 0.8868, Adjusted R-squared: 0.882

F-statistic: 187.9 on 1 and 24 DF, p-value: 7.576e-13

**The estimated model**

**exam = -13.9 + 1.15 testave +** , where  is N(0,8.0)

**So, on average, a student’s exam % is**

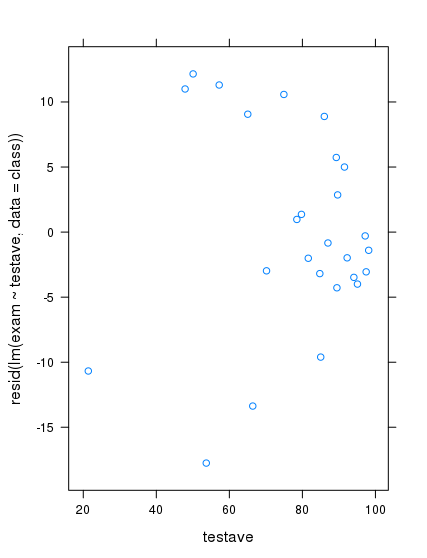
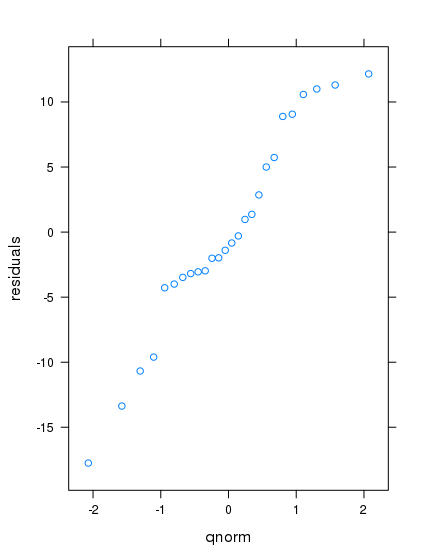
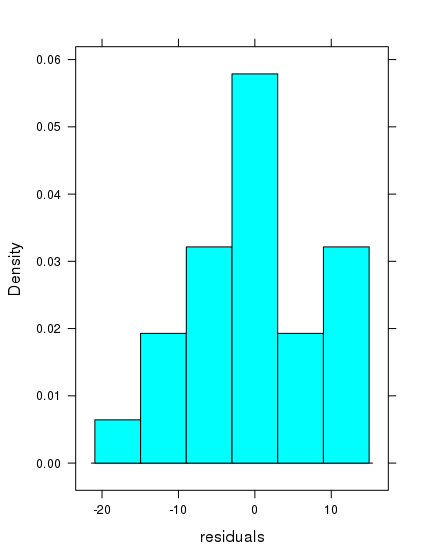
**Check Model Assumptions**

> residuals<-resid(lm(exam~testave,data=class))

> histogram(residuals)

> qqmath(residuals)

> xyplot(resid(lm(exam~testave,data=class))~testave,data=class)



**Are the residuals normally distributed?**

**From the scatter plot and the qq-plot, normality isn’t perfect, but not too bad for a relatively small sample.**

**Is the amount of variability in the residuals independent of x?**

**As testave gets closer to 100, the variability in exam shrinks. Is that surprising?**

**95% CI for the average of all students with a test average = 85%**

> mod<-makeFun(lm(exam~testave,data=class))

> mod(85,interval="confidence",level=.95)

fit lwr upr

1 83.82115 80.36521 87.27709

**Todd has a test average of 85%. How well do you expect him to do on the final exam? Express this as a 95% PI.**

> mod(85,interval="prediction",level=.95)

fit lwr upr

1 83.82115 67.01541 100.6269

**CAVEAT** : The CI and PI interval results presume that the underlying model is a simple linear regression model; i.e., the term is N(0,sigma) where sigma is independent of x. So, if the check model assumptions graphs are bad enough, you cannot have a lot of faith in the CI and PI. In our case, the graphs are not great, but not “terrible” (IMO).

**Exercises 15**

**LOAD THE PACKAGE. Lock5withR.**

The data frame **StudentSurvey** contains data on 362 college students. Below is a list of the variables included;

* Year Year in school
* Gender Student's gender: F or M
* Smoke Smoker? No or Yes
* Award Preferred award: Academy Nobel Olympic
* HigherSAT Which SAT is higher?  Math or Verbal
* Exercise Hours of exercise per week
* TV Hours of TV viewing per week
* Height Height (in inches)
* Weight Weight (in pounds)
* Siblings Number of siblings
* BirthOrder Birth order, 1=oldest
* VerbalSAT Verbal SAT score
* MathSAT Math SAT score
* SAT Combined Verbal + Math SAT
* GPA College grade point average
* Pulse Pulse rate (beats per minute)
* Piercings Number of body piercings

1.Use the **head** command to produce the first 6 lines in the data-frame: StudentSurvey. Let **VerbalSAT** be the explanatory variable and **MathSAT** be the response variable.

a. Create the scatter plot.

b. Apply the summary command to get the estimates for the three model parameters , , and . Write out the estimated linear regression model using these estimates.

c. How good is VerbalSAT as a predictor of MathSAT; i. e., what percent of the variation in MathSAT scores is explained by the VerbalSAT scores? (Remember, that is the R-square value)

d. Check the normality assumption by plotting a histogram of the residuals and a qqmath plot of the residuals. Include the plots and your judgment about whether the normality assumption holds as your solution to (d).

e. Check the equal variance assumption by plotting the residuals vs verbalSAT. Include the plot and your judgment about whether the equal variance assumption holds as your solution to (e).

f. Use the simple linear regression model to estimate the average MathSAT score for all students with VerbalSAT = 650 using a 95% CI.

g. Linda comes to you and tells you that she has a VerbalSAT score of 700 and asks you to guess her MathSAT score. What would your guess be as a 95 PI?

2. Repeat Exercise (1) using **MathSAT** as the explanatory variable and **GPA** as the response variable. (For parts (f) and (g) use MathSAT = 650 and MathSAT = 700)