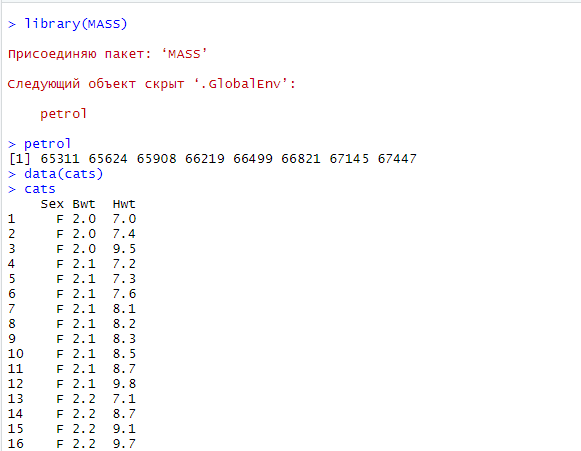
**5.2 Simplelinearregression,prediction: Heartandbodyweights?**

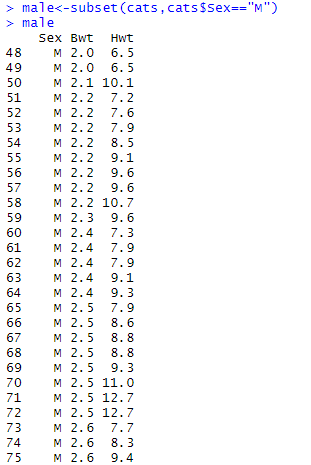
1. In the R package MASS there is a dataset called cats. Run the following commands:

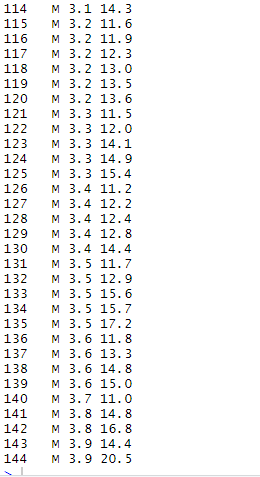
library(MASS) data(cats)



Have a look at the dataset. The variables Bwt and Hwt give the weight of the body (kg) and the heart (g), respectively. There are both male and female cats. Make a dataset with the data from males only.

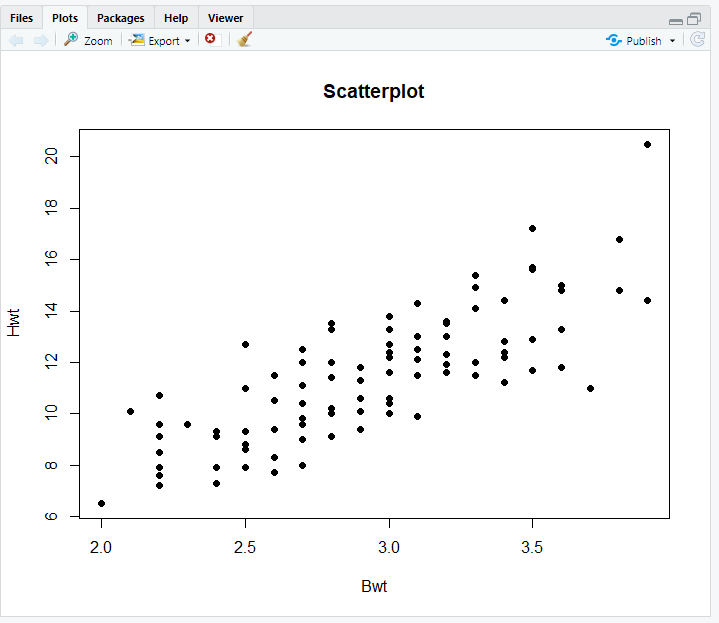
**male<-subset(cats,cats$Sex=="M")**





2. Make a scatterplot of the data for the male cats (Bwt on x-axis, Hwt on y-axis). Does it look reasonable to use a linear regression model for the data?

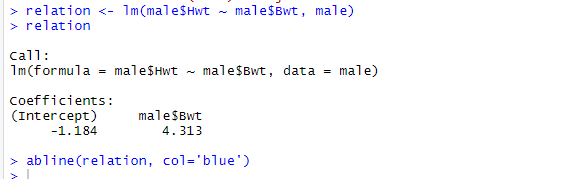
**plot(male$Bwt, male$Hwt, main="Scatterplot",xlab="Bwt", ylab="Hwt", pch=20)**

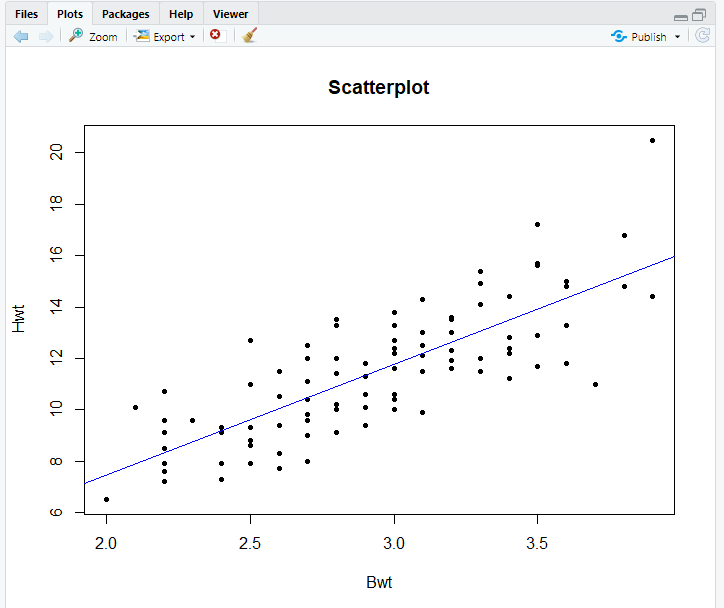


3. Fit a linear regresison model for the male cats, that allows for prediction of the heart weight given the body weight. Add the ﬁtted regression line to the scatterplot from the previous question.

**relation <- lm(male$Hwt ~ male$Bwt, male)**

**abline(relation, col='blue')**





4. Find the coefﬁcients of the ﬁtted line. How large is the expected difference in heart weight for two cats with a difference of 1 kg in bodyweight? Find a conﬁdence interval for this difference? How large is the expected difference in heart weight for two cats with a difference of 100 g in bodyweight?

**summary(relation)**

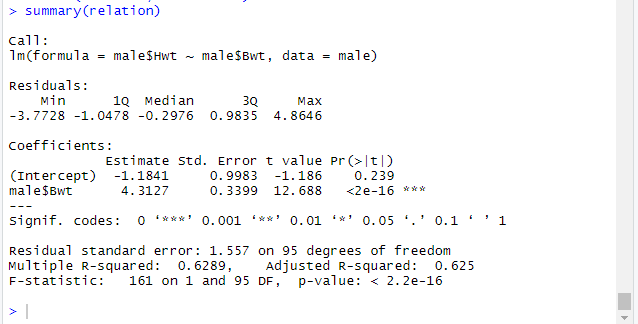
**ss<-subset(male,diff(male$Hwt)==1)**

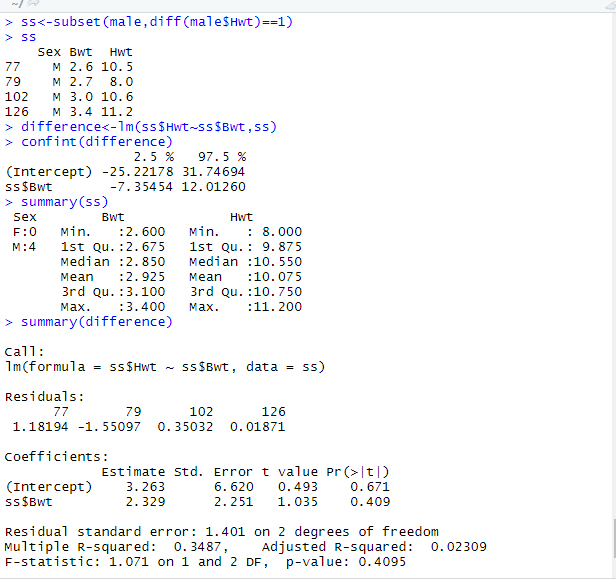
**difference<-lm(ss$Hwt~ss$Bwt,ss)**

**confint(difference)**

**summary(ss)**

**summary(difference)**





5. Use model validation plot to examine if the model is appropriate for the data.

**plot(fitted(relation), residuals(relation), col='green')**

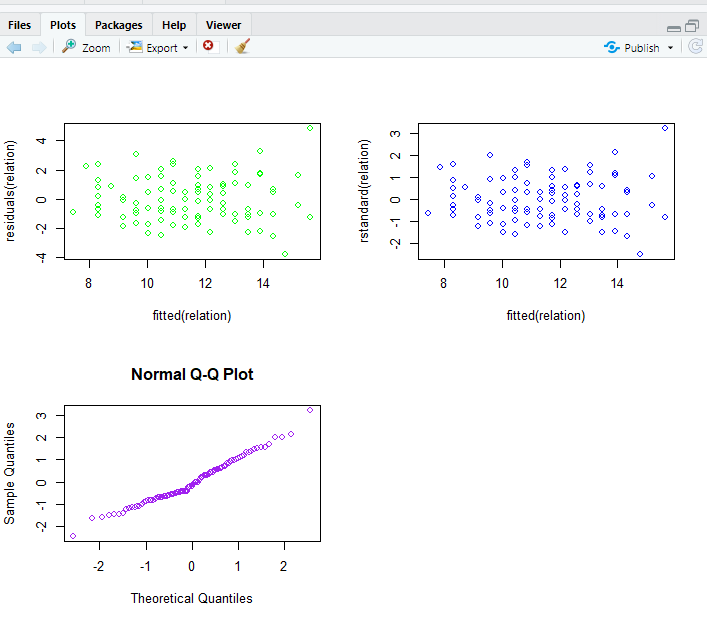
**plot(fitted(relation), rstandard(relation), col='blue')**

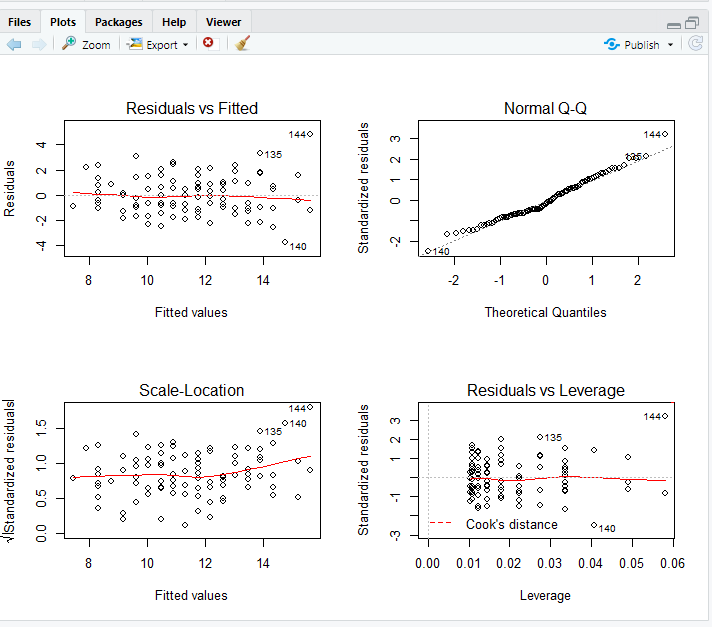
**qqnorm(rstandard(relation), col='purple')**

**abline(0,1)**

**par(mfrow=c(2,2))**

**plot(relation)**

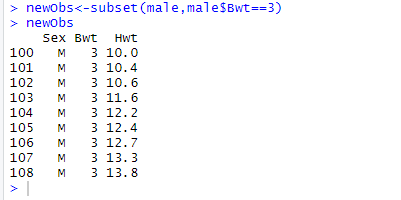




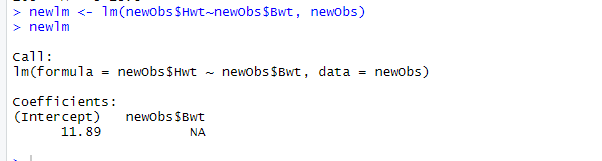
6. Use the estimates to ﬁnd the expected heart weight for a male cat that weighs 3 kg. Then try the commands (where you replace the name regModel with whatever name you gave the the model ﬁt in question 2).

newObs <- data.frame(Bwt=3) newObs predict(regModel, newObs) predict(regModel, newObs, interval="predict")

**newObs<-subset(male,male$Bwt==3)**

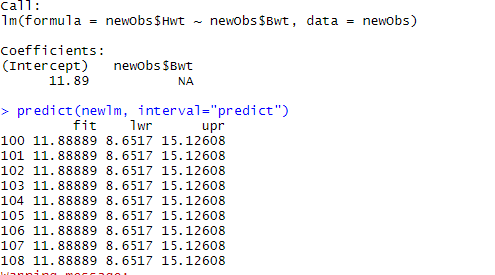


**newlm <- lm(newObs$Hwt~newObs$Bwt, newObs)**



**lm(formula = newObs$Hwt ~ newObs$Bwt, data = newObs)**

**predict(newlm, interval="predict")**



**5.3 Multiplelinearregression: Toxicityofdissolutions?**

Prerequisites: Exercises 3.1, 3.2 (reading and working with data) and 5.1 (linear regression) Data from 24 chemical dissolutions have been collected in order to examine the association between the toxicity of the dissolution on the one side and three explanatory variables on the other side. The data are saved in the ﬁles lser.xlsx and lser.csv with the following variables:

• tox: Toxicity of the dissolution • base: Ablility to accept hydrogen ions

• acid: Ablility to liberate hydrogen ions • colour: Ability to change colour 1. Make an R dataset called lser with the data, and use the command plot(lser) to get an overview of the data.

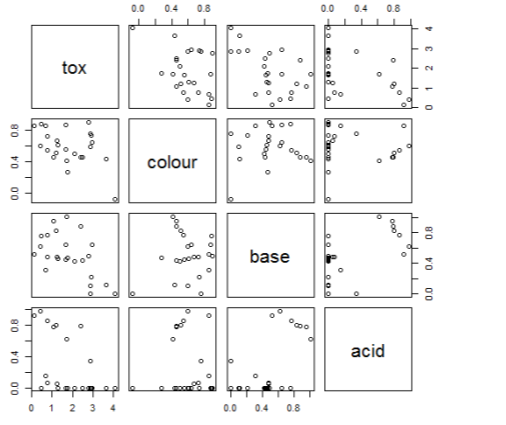
**getwd()**

**lser <- read.csv("Iser.csv")**

**library(xlsx)**

**lser<-read.xlsx("Iser.xlsx",1)**

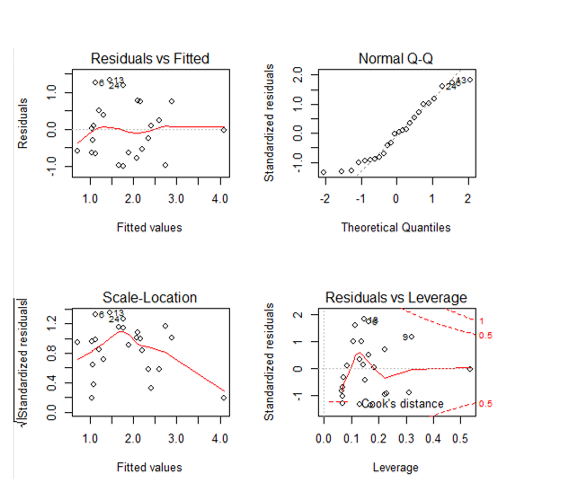
**plot(lser)**

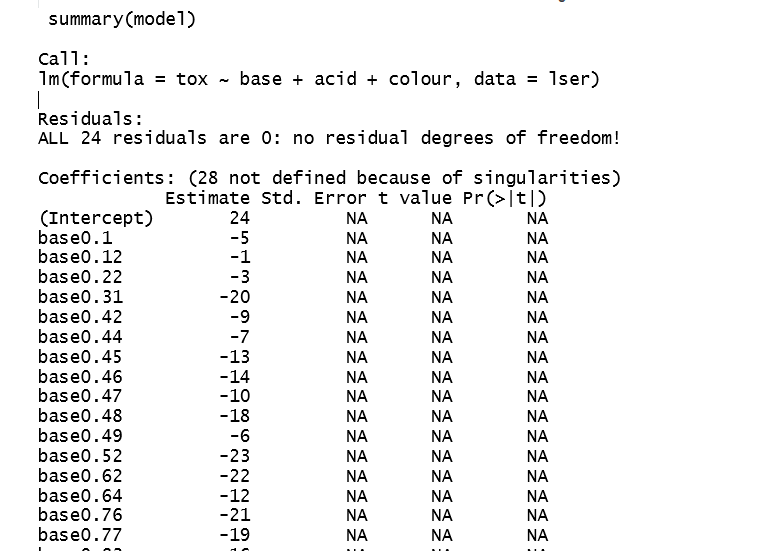


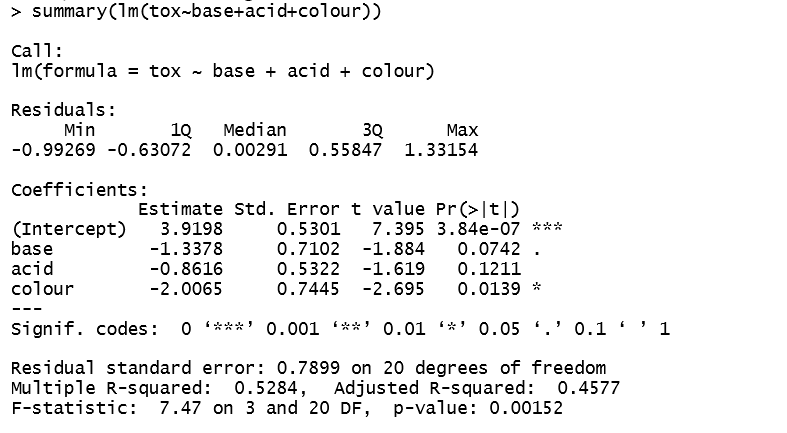
2. Fitamultiplelinearregressionmodelwith tox asresponseand base, acid and colour as explanatory variables. Is the model appropriate for the data? What is the interpretation of the parameter estimates?

**mod<-lm(tox~base+acid+colour, lser)**

**plot(mod)**

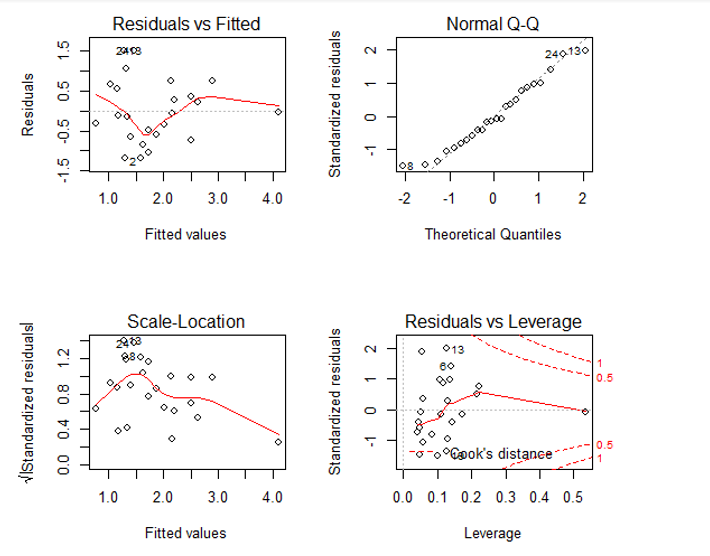




3. Are all three explanatory variables signiﬁcant? Remove insigniﬁcant variables (one at a time) until all terms are 

**model\_no\_sig <- lm(tox~base+colour, data=lser)**

**plot(model\_no\_sig)**



signiﬁcant.

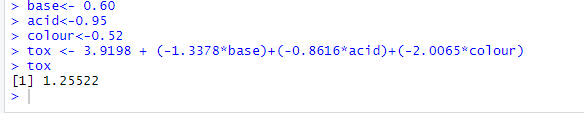
4. Calculate the expected toxicity for a solvent which has base=0.60, acid=0.95, and colour=0.52.

**base<- 0.60**

**acid<-0.95**

**colour<-0.52**

**tox <- 3.9198 + (-1.3378\*base)+(-0.8616\*acid)+(-2.0065\*colour)**



5.5 Logisticregression: Pneumoconiosisamongcoalminers

Prerequisites: None (although Exercise 5.1 is perhaps useful for understanding the output) Binary variables arevariables with two possible outcomes, for example dead or alive, healthy orsick,germinatedornot. Logisticregressionisrelevantwhentheresponsevariableisbinary. Theaimistorelatetheprobabilitiesofthetwooutcomestooneormoreexplanatoryvariables. In the simplest case with one continuous covariate, x, and binary response, y, the logistic regression model assumes that log-odds is a linear function of x. Let p=P(y=1), and recall that the odds is deﬁned as P(y = 1)/P(y = 0) = p/(1−p). Then the assumption is log\_ p 1−p\_=α+βx. (3) Just as in ordinary linear regression (with continuous response) the ﬁrst aim is to estimate the parametersα andβ. This exercise is about pneumoconiosis among coalface workers. Data were collected in order to examine the relationship between exposure time (years) and risk of disease. Severity of disease was originally rated into three categories, but here we will use only two (normal and diseased):

Exposure time Normal Diseased 5.8 98 0 15 51 3 21.5 34 9 27.5 35 13 33.5 32 19 39.5 23 15 46 12 16 51.5 4 7

The data are saved in the ﬁles coalworker1.xlsx and coalworker1.csv. Notice how, for each person, it has been observed whether he was diseased or not. This correspondstoabinaryvariable. Inthefollowingwewillconsideramodelwithlog-exposure as covariate, i.e. with log(exposure) as the x-variable in (3).

1. Read the data into R; call the dataset coalworker1. Then try the following commands and make sure you understand what happens:

total <- normal + diseased relativeFreq <- diseased/total logOdds <- log(relativeFreq / (1-relativeFreq)) plot(log(exposure), logOdds)

**coalworker1<-read.csv("coalworker1.csv", dec=',', sep=';')**

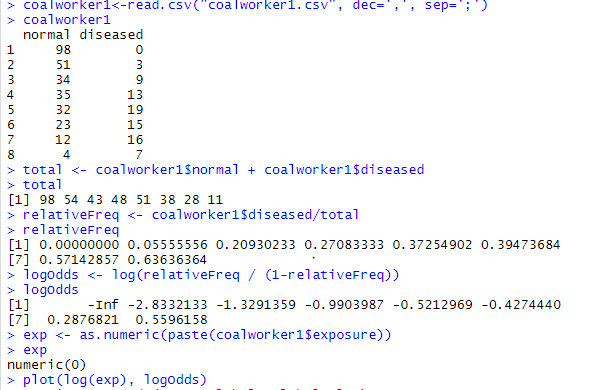
**total <- coalworker1$normal + coalworker1$diseased**

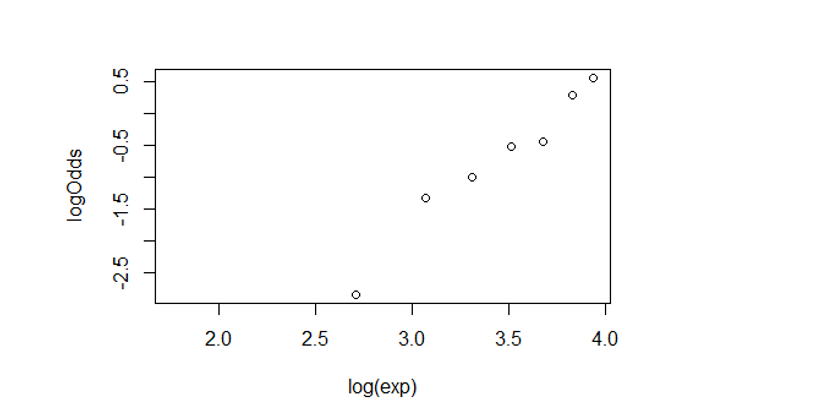
**relativeFreq <- coalworker1$diseased/total**

**logOdds <- log(relativeFreq / (1-relativeFreq))**

**exp <- as.numeric(paste(coalworker1$exposure))**

**plot(log(exp), logOdds)**

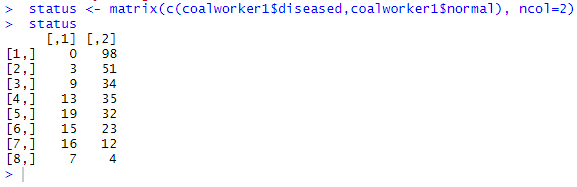




What does the plot tell you about the appropriateness of the logistic regression model with log-exposure as explanatory variable? 2. Fit the logistic regression model with the following commands:

status <- matrix(c(diseased,normal), ncol=2) status logreg1 <- glm(status~log(exposure), family=binomial)

Noticetheoption family=binomial. Ittells R tointerpretvaluesinthematrix status as outcomes from a binomial distribution.

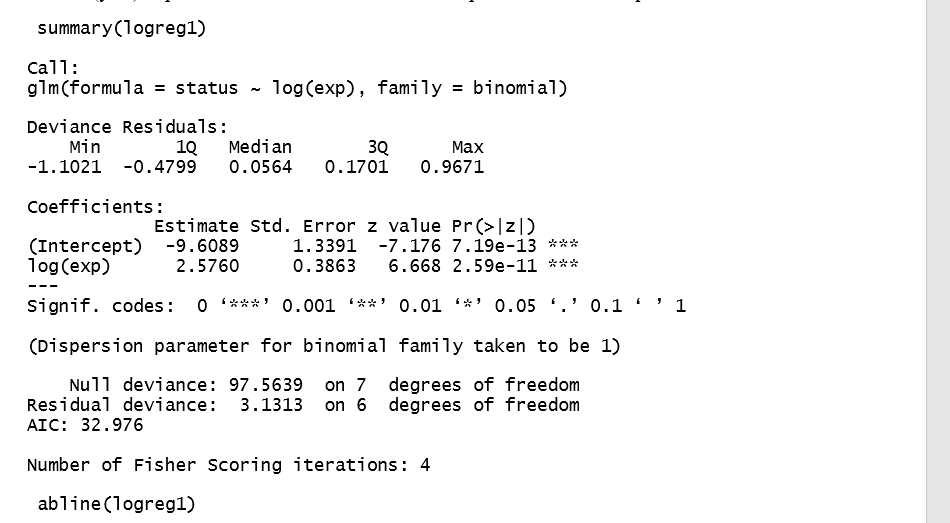


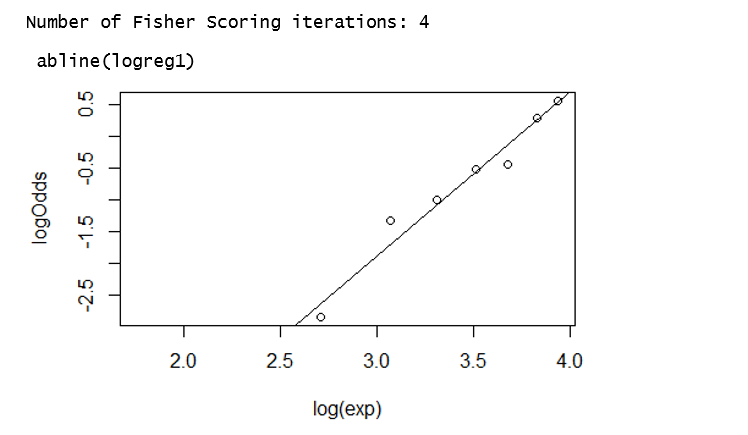
3. Try the commands

summary(logreg1) abline(logreg1)

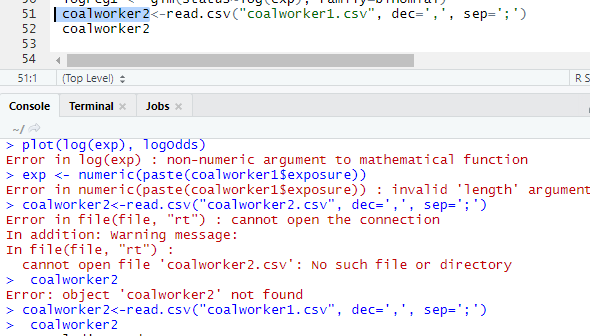
What are the estimated values of α and β? Does the model seem to ﬁt the data reasonably well?

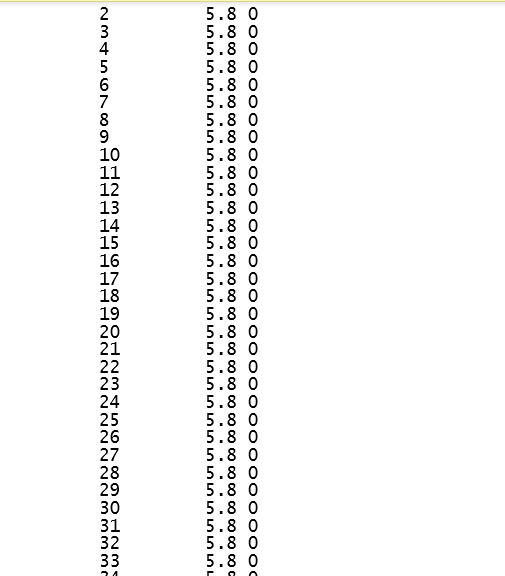
Above, the data was represented by the numbers of diseased and normal for each level of exposurewith8observationsintotal. Thedatacouldalsoberepresentedwithoneobservation per person (371 observations in total). The ﬁles coalworker2.xlsx and coalworker2.csv contain the data in this representation with the variable y telling whether the person was diseased (y=1) or normal (y=0). Apart from this there is as variable exposure2 with the exposure time.

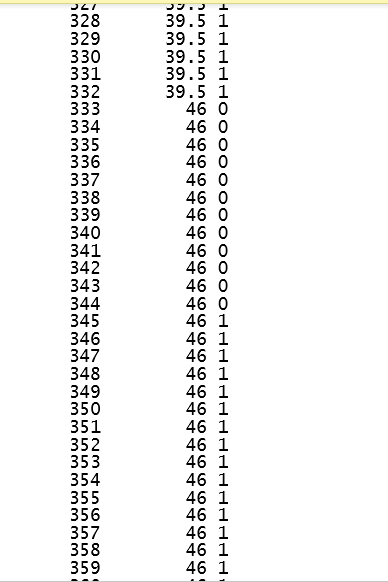




4. Read the data into R in the new form; call the dataset coalworker2. Make sure you understand the structure of the new dataset, and make sure you understand that coalworker1 and coalworker2 contain the same information.







5. When the data is represented as in coalworker2, the logistic regression model is ﬁtted as follows:

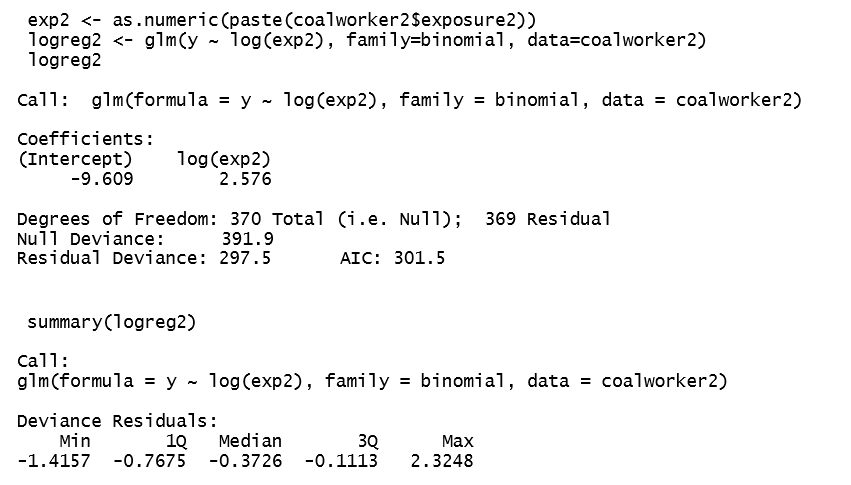
logreg2 <- glm(y ~ log(exposure2), family=binomial, data=coalworker2) summary(logreg2)

Fit the model and make sure that you get the same estimates as you did with logreg1 before.

The last questions are about the interpretation of the estimates and the models.

**exp2 <- as.numeric(paste(coalworker2$exposure2))**

**logreg2 <- glm(y ~ log(exp2), family=binomial, data=coalworker2)**



6. Consider a person with exposure time equal to 30 years. What is the estimated logodds of this person being diseased? What is the estimated probability that the person is diseased? Hints: For the estimation of log-odds, remember that the explanatory variable is the log-transformed exposure time. For the probability you sh

**newCoalWorker <- subset(coalworker1, coalworker1$exposure == 33.5)**

**newCoalWorker**

**total <- newCoalWorker$normal + newCoalWorker$diseased**

**relativeFreq <- newCoalWorker$diseased/total**

**logOdds1 <- log(relativeFreq / (1-relativeFreq))**

**newCoalWorker1 <- subset(coalworker1, coalworker1$exposure == 27.5)**

**newCoalWorker1**

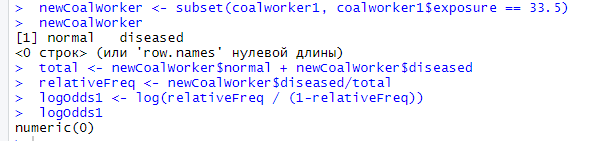
**total <- newCoalWorker1$normal + newCoalWorker1$diseased**

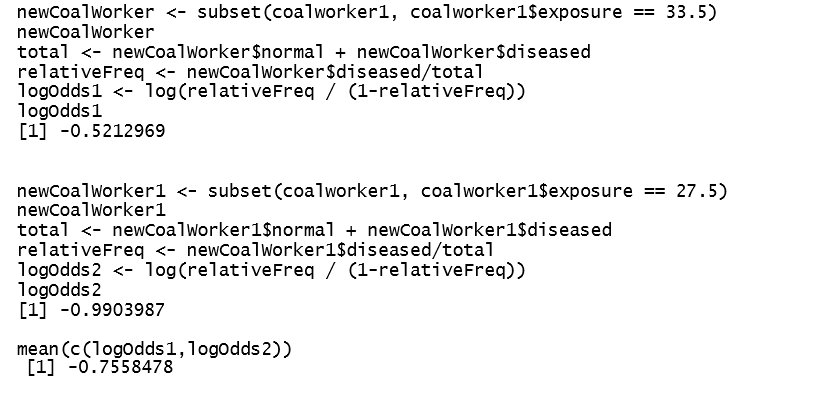
**relativeFreq <- newCoalWorker1$diseased/total**

**logOdds2 <- log(relativeFreq / (1-relativeFreq))**

**logOdds2**

**mean(c(logOdds1,logOdds2))**



ould solve expression (3) for p.

7. How many years of exposure gives a 50% risk of having developed the disease? Hint: We are looking for the exposure time corresponding to p = 0.5. What is the correspondingvalueoflog-odds? Usetheestimatestocomputethe(logarithmic)exposure time.

**total\_diseased <- sum(coalworker1$diseased)**

**p<-0.5**

**total\_diseased\*p**

**mean(coalworker1$diseased)**

**newcoal <- subset(coalworker1, coalworker1$diseased == 19 | coalworker1$diseased == 15 | coalworker1$diseased == 7)**

**exp <- as.numeric(paste(newcoal$exposure))**

**sumExp <- sum(exp)**

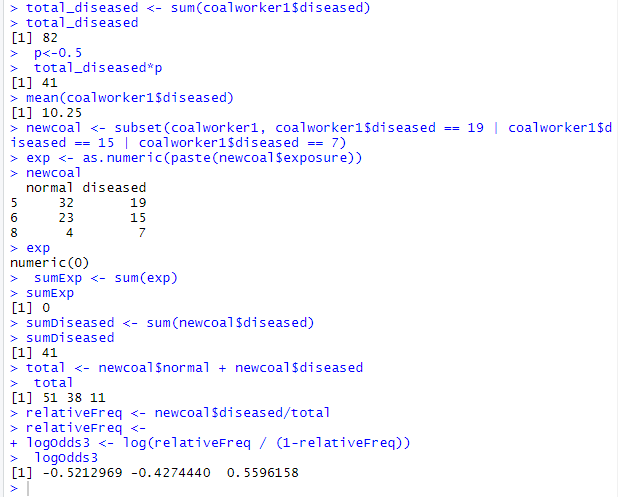
**sumDiseased <- sum(newcoal$diseased)**

**total <- newcoal$normal + newcoal$diseased**

**relativeFreq <- newcoal$diseased/total**

**logOdds3 <- log(relativeFreq / (1-relativeFreq))**

**logOdds3**



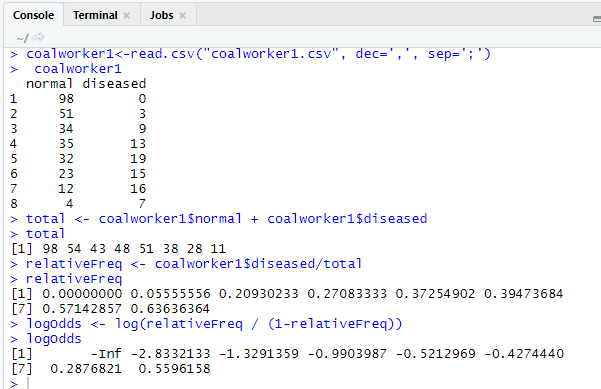
8. What happens with the odds if the exposure time is doubled? Hint: What happens to log-exposure if exposure time is doubled? What is the effect on log-odds? Which effect on odds does that correpsond to?

**coalworker1<-read.csv("coalworker1.csv", dec=',', sep=';')**

**total <- coalworker1$normal + coalworker1$diseased**

**relativeFreq <- coalworker1$diseased/total**

**logOdds <- log(relativeFreq / (1-relativeFreq))**



**When exposure time is doubled, the function log odds will be doubled the output too.**