Test 2 R Script

Alexander Hernandez

11/22/2022

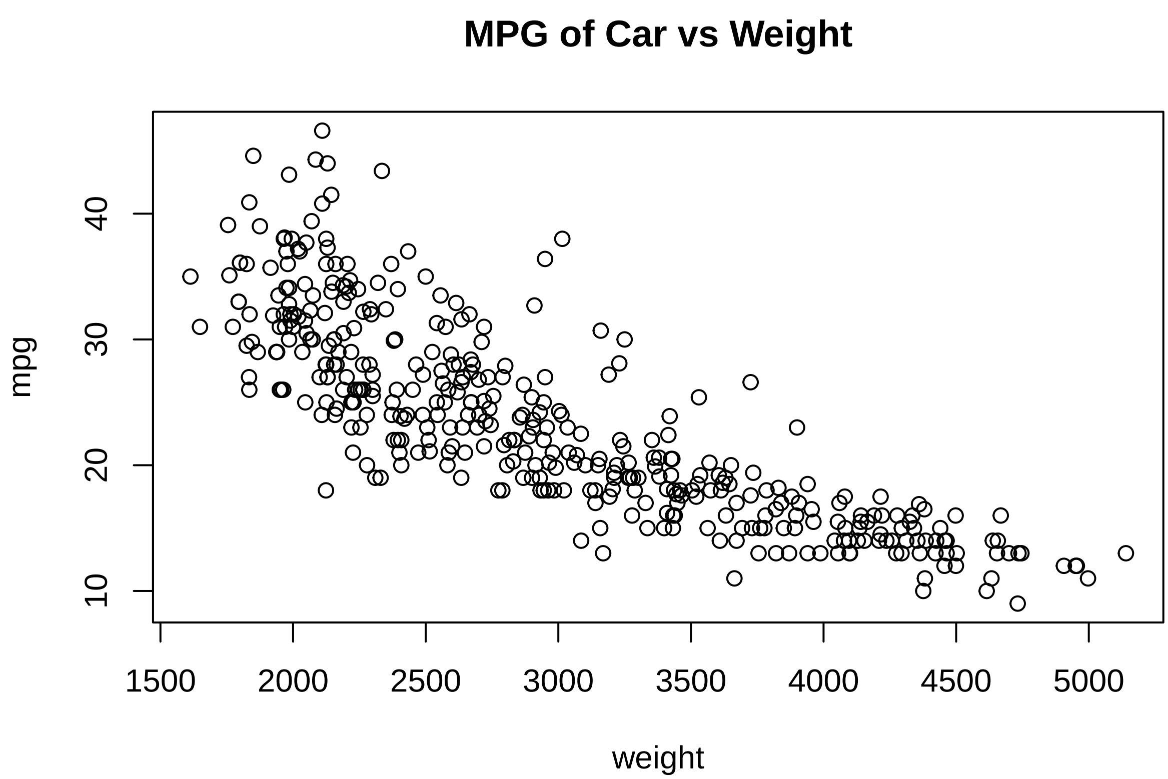
```
library(MASS)
```

1) Fuel Consumption by Weight

a) Import the data in R and display with scatterplot

```
fuel = read.csv("http://www.stat.cmu.edu/~cshalizi/mreg/15/hw/04/auto-mpg.csv")
attach(fuel)

plot(mpg ~ weight,
     main="MPG of Car vs Weight")
```



b) Fit a simple linear regression model and state equation. Provide interpretation of parameter B1 to determine the relationship between weight and fuel consumption.

```
model1 = lm(mpg ~ weight)
summary(model1)
```

```
##
## Call:
## lm(formula = mpg ~ weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.012  -2.801  -0.351   2.114  16.480
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.3173644  0.7952452   58.24   <2e-16 ***
## weight      -0.0076766  0.0002575  -29.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.345 on 396 degrees of freedom
## Multiple R-squared:  0.6918, Adjusted R-squared:  0.691
## F-statistic: 888.9 on 1 and 396 DF,  p-value: < 2.2e-16
```

```
#   y =        B0 +       B1(x)
# mpg = 46.317364 - 0.007677(weight)
# As B1 (weight) increases, the mpg decreases.
```

c) Determine the coefficient of determination of the model and provide its interpretation.

```
# According to the summary, the R**2 is 0.6918.
```

d) Use the model to predict the mpg if the car is 2100 lbs. 90% conf interval

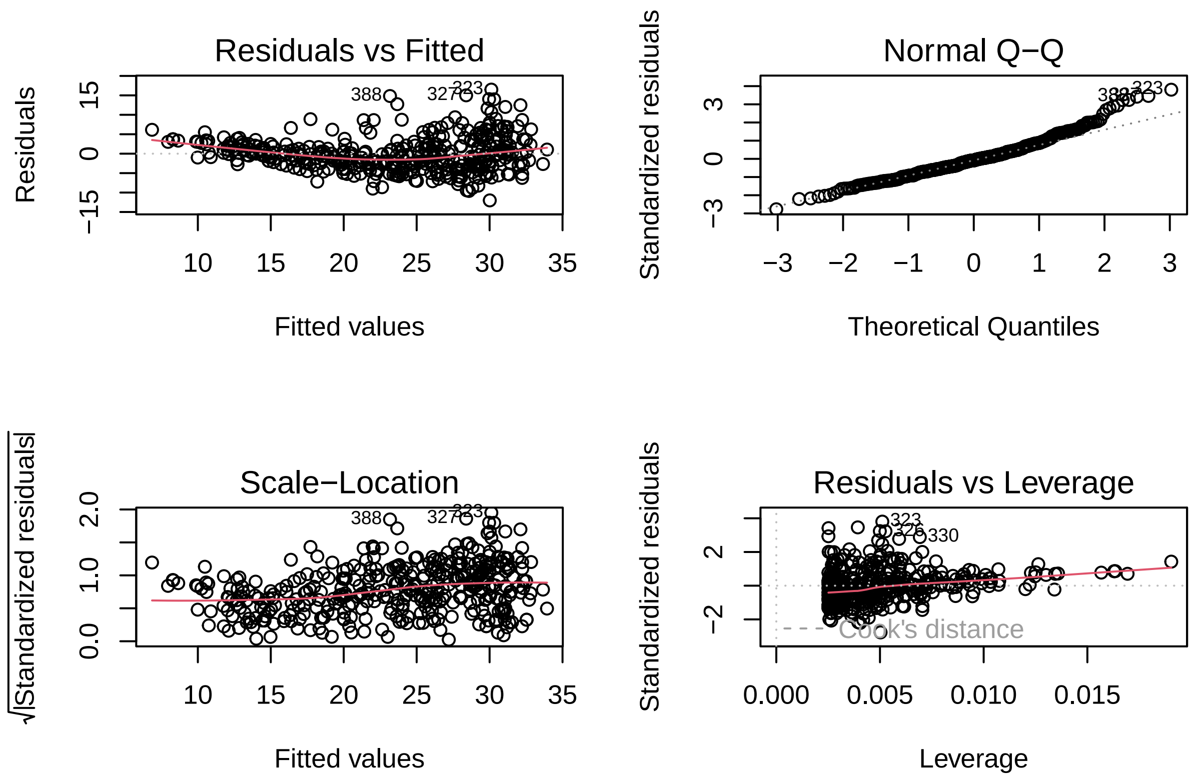
```
predict(model1, data.frame(weight=2100), interval="conf", level=0.95)
```

```
##        fit      lwr      upr
## 1 30.19648 29.58211 30.81085
```

```
# Predicted MPG: 30.19648
# Conf Interval: (29.58211, 30.81085)
```

e) Perform the residual analysis of the model

```
par(mfrow=c(2,2))
plot(model1)
```




```
# The residual plots seem valid enough, but they could be better.
```

2) Loblolly

a) Extract the variable names and dimensions of the data

```
names(Loblolly)
```

```
## [1] "height" "age"    "Seed"
```

```
dim(Loblolly)
```

```
## [1] 84  3
```

```
# There are 84 observations with 3 variables, "height, "age", and "seed".
```

b)Does the relationship between age and height of the tree appear linear? If so, please determine the linear model and display with scatterplot

```
attach(Loblolly)
plot(height ~ age,
     main="Height of Loblolly Pine Trees vs Age")
# The relationship appears linear

model2b= lm(height ~ age)
model2b
```

```
##
## Call:
## lm(formula = height ~ age)
##
## Coefficients:
## (Intercept)          age
##      -1.312        2.591
```

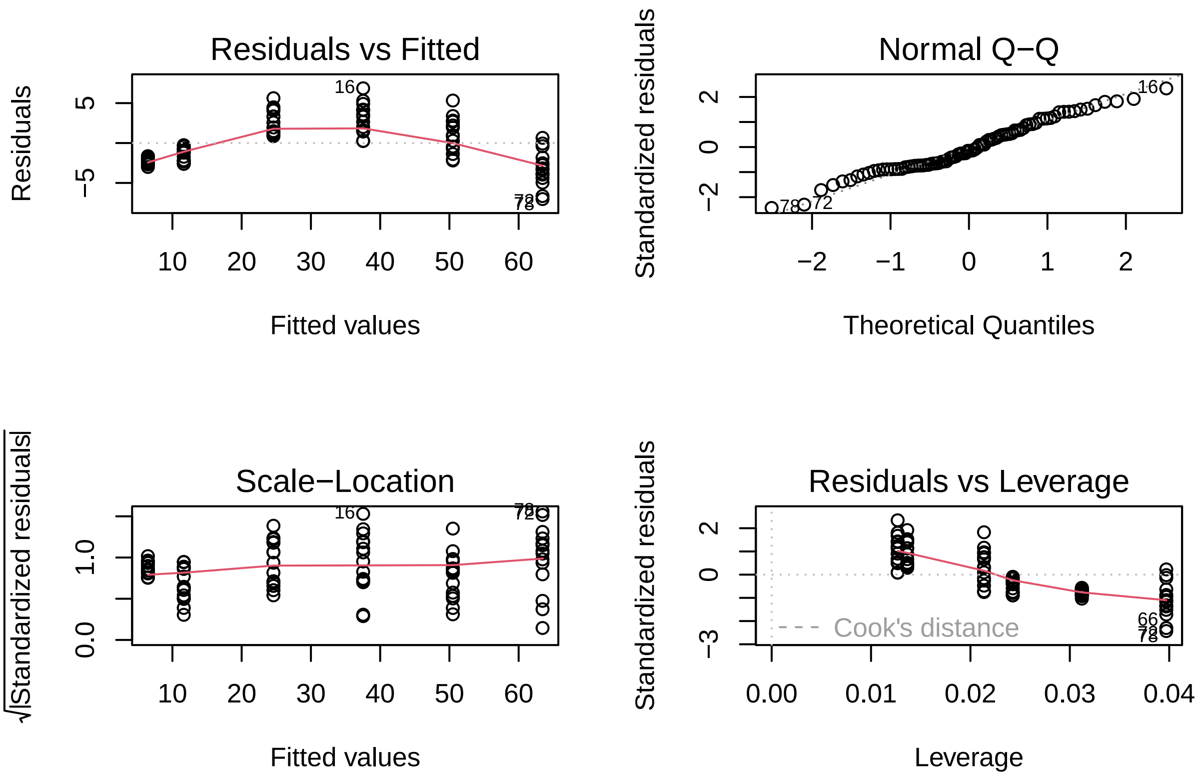
```
#      y =     B0 +    B1(x)
# height = -1.312 + 2.591(age)

abline(model2b, col=2)
```

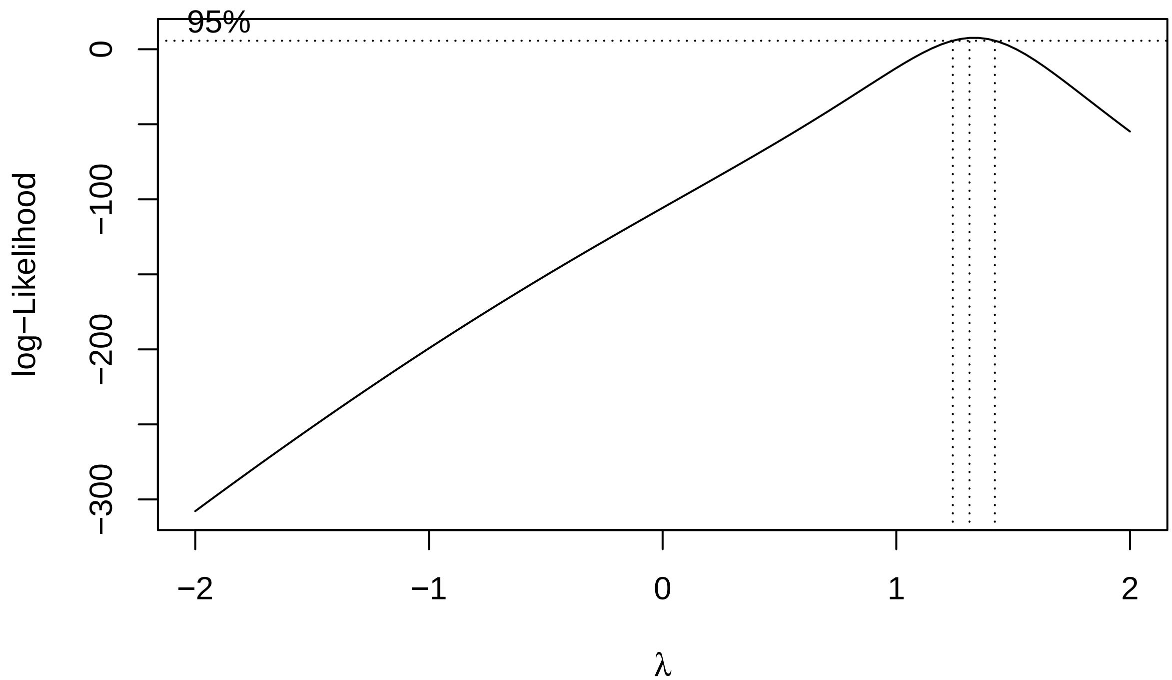
Height of Loblolly Pine Trees vs Age
height
60
50
40
30
20
10
5
10
15
20
25
age

c) Perform the residual analysis to check whether a transformation is needed. If so, what is the appropriate value of the transformations?

```
par(mfrow=c(2,2))
plot(model2b)
```




```
par(mfrow=c(1,1))
boxcox(model2b)
```

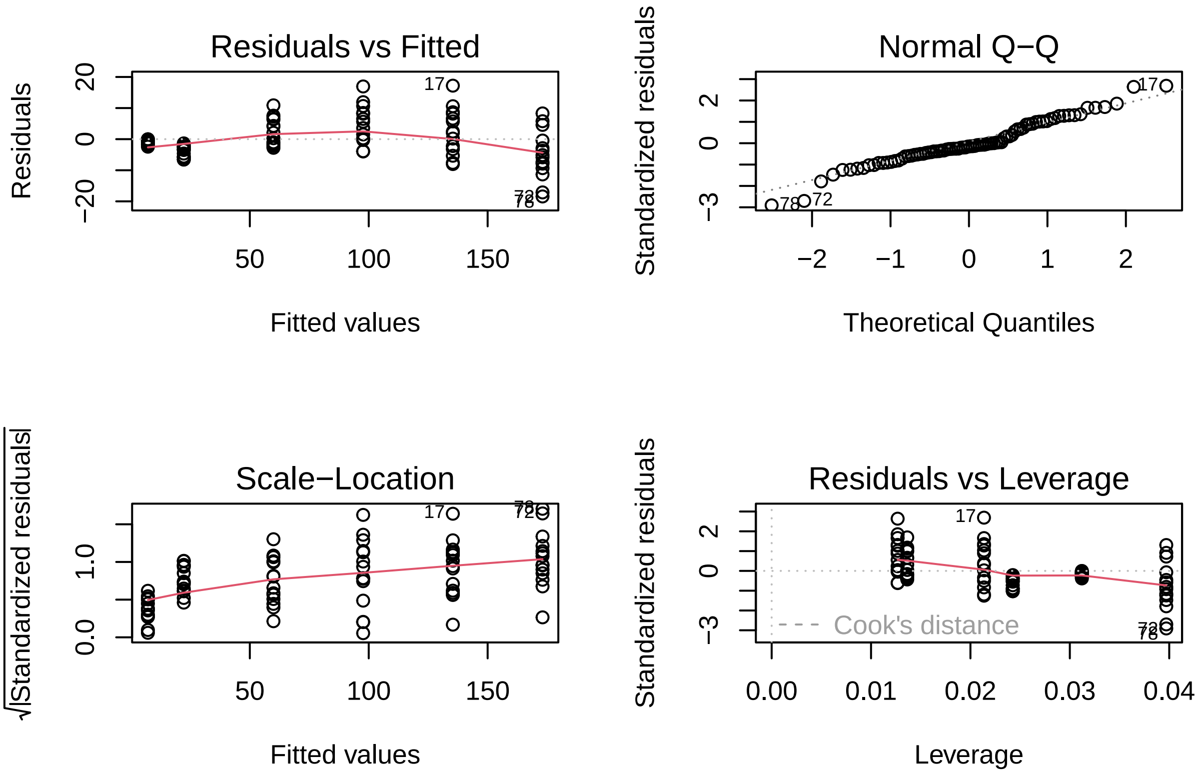



```
# Lambda value of 1.25 may be useful
```

d) Is the transformation worth it?

```
model2c = lm(height**1.25 ~ age)

par(mfrow=c(2,2))
plot(model2c)
```




```
# Residuals are more balanced so,
# the transformation slightly improved the model.
```

3) Leukemia Remission

a) Import the data to determine how many remission cases of leukemia are in the dataset

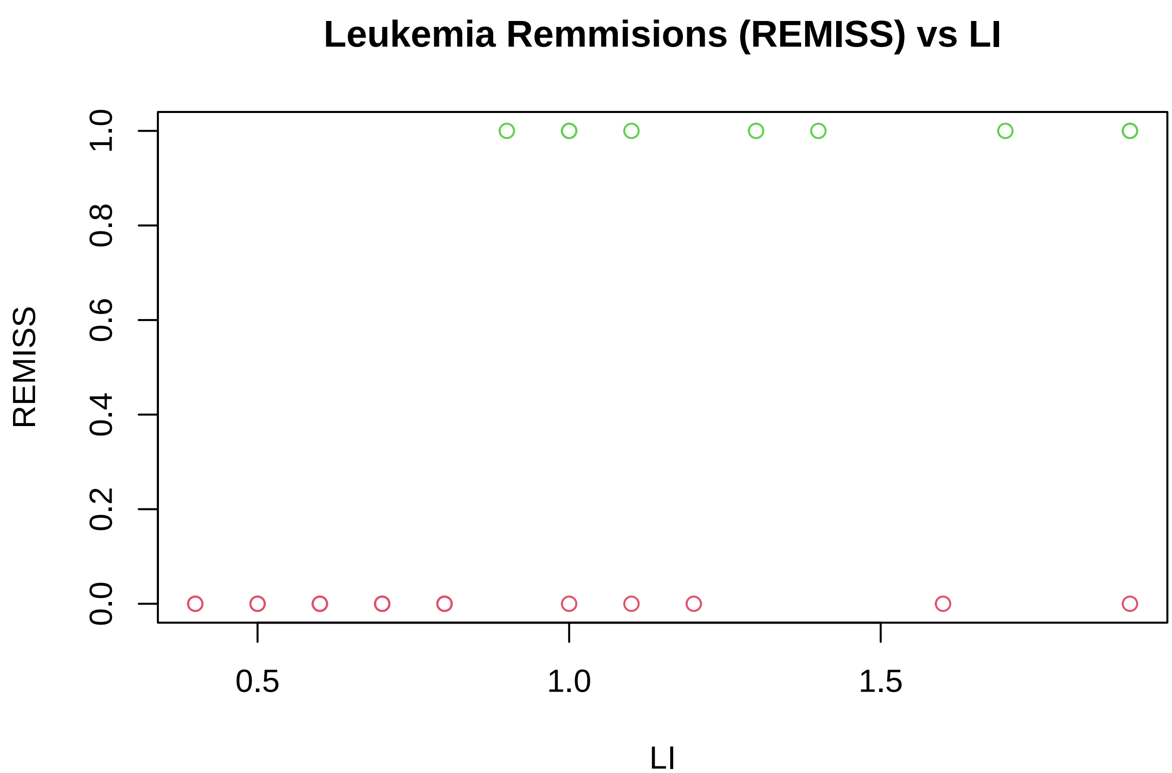
```
leukemia = read.table("C:\\repos\\STAT 50001\\Test 2\\leukemia.txt",
                      header=TRUE)
attach(leukemia)

sum(leukemia["REMISS"])
```

```
## [1] 9
```

b) Display the variable REMISS as a response variable using LI as a predictor variable

```
par(mfrow=c(1,1))
plot(REMISS ~ LI, main="Leukemia Remmisions (REMISS) vs LI",
     col=ifelse(leukemia["REMISS"] == 1, 3, 2))
```



c) Fit a simple logistic regression model and write the equation of the model

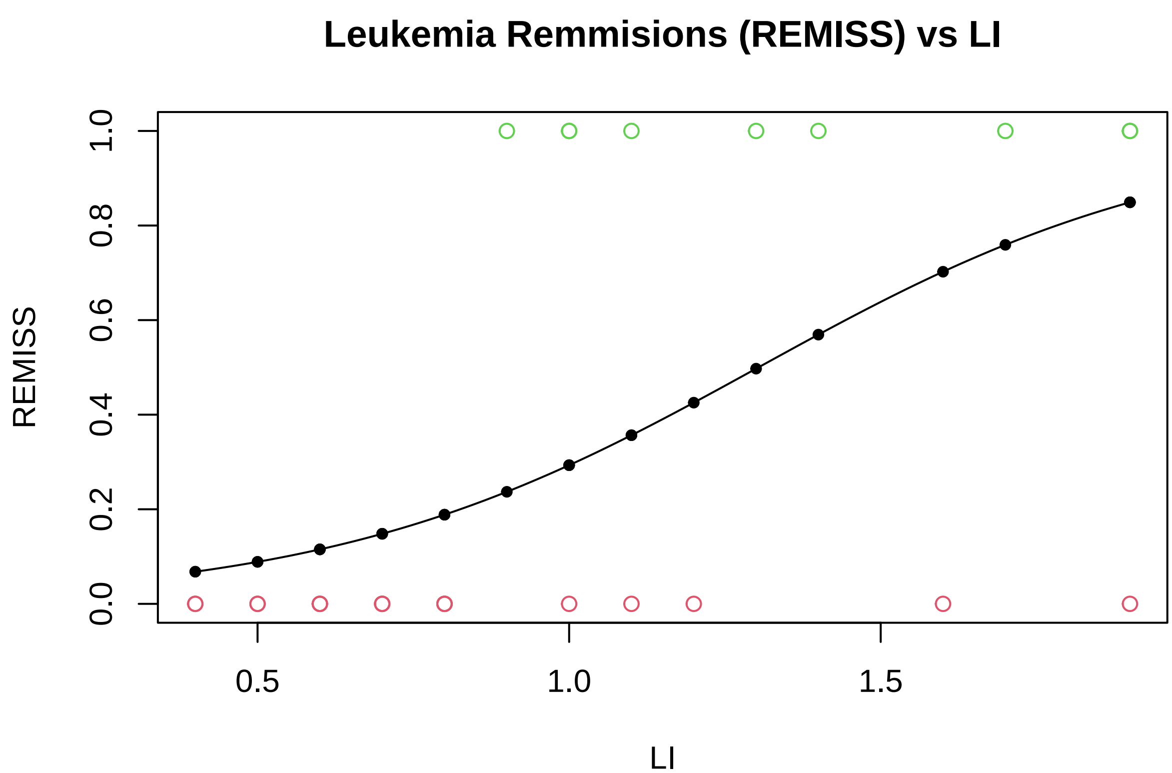
```
model3 = glm(REMISS ~ LI,
             family = binomial(logit))
model3
```

```
##
## Call:  glm(formula = REMISS ~ LI, family = binomial(logit))
##
## Coefficients:
## (Intercept)           LI
##      -3.777        2.897
##
## Degrees of Freedom: 26 Total (i.e. Null);  25 Residual
## Null Deviance:       34.37
## Residual Deviance: 26.07     AIC: 30.07
```

```r
#   E(y) = [1 + exp(  -B0 -   B1(x)   )]^-1
# REMISS = [1 + exp( 3.77 - 2.897(LI) )]^-1
```

d) Display the probability curve along with the scatterplot

```r
plot(REMISS ~ LI, main="Leukemia Remmisions (REMISS) vs LI",
     col=ifelse(leukemia["REMISS"] == 1, 3, 2))
curve(predict(model3, data.frame(LI=x), type="resp"), add=TRUE)
points(LI, fitted(model3), pch=20)
```



e) Calculate the probability Leukemia Remission if percentage labeling index of the bone marrow leukemia cells (LI) is 1.7

```
predict(model3, data.frame(LI=1.7), type="resp")
```

```
##         1
## 0.7591835
```

```
# 0.7591835 REMISS probability if LI=1.7
```

4) Effect of Drug in Reduction of Excess Body Weight

a) Fit a multiple linear regression model reflecting the effect of gender

```
drug = read.table("C:\\repos\\STAT 50001\\Test 2\\drug.txt",
                  header=TRUE)
attach(drug)
```

```
## The following object is masked from Loblolly:
##
##     age
```

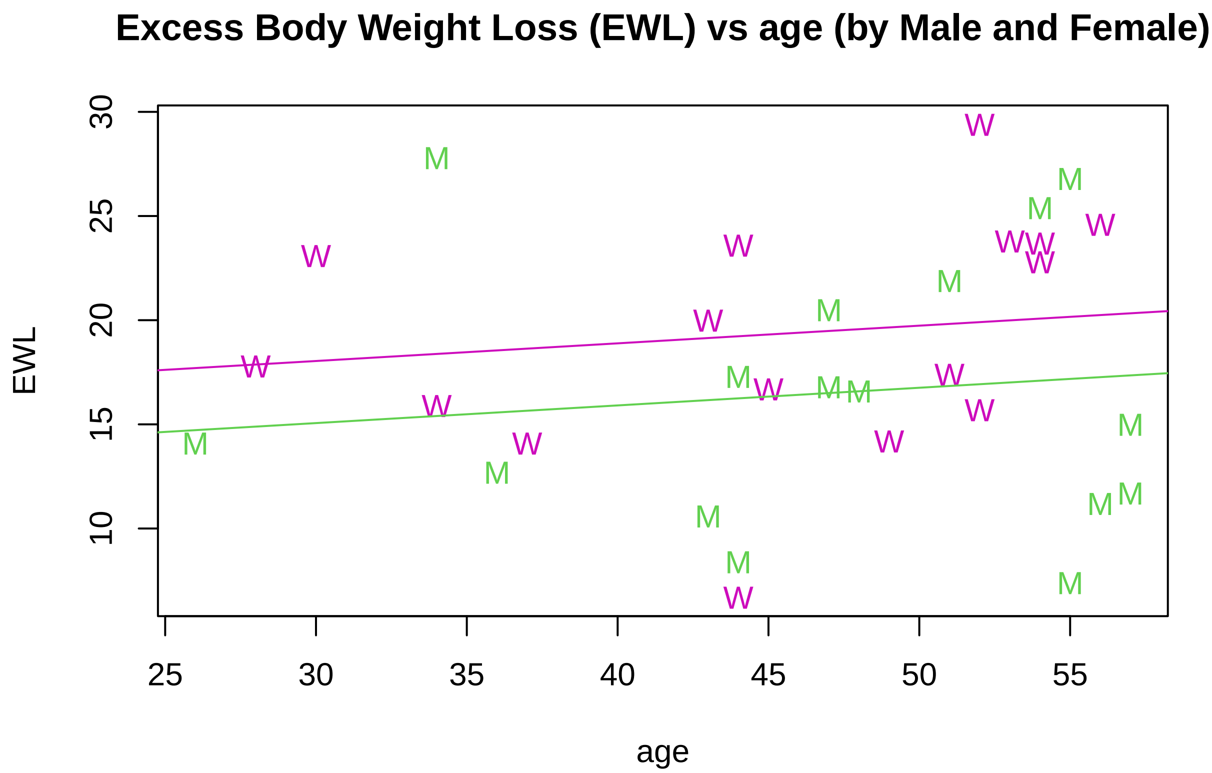
```
model4 = lm(EWL ~ age + gender)
summary(model4)
```

```
##
## Call:
## lm(formula = EWL ~ age + gender)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5271  -4.2876  -0.0284   4.0873  12.4007
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.49515    5.76280   2.689   0.0118 *
## age          0.08482    0.12255   0.692   0.4944
## gender      -2.97968    2.15070  -1.385   0.1765
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.053 on 29 degrees of freedom
## Multiple R-squared:  0.07139,    Adjusted R-squared:  0.007347
## F-statistic: 1.115 on 2 and 29 DF,  p-value: 0.3417
```

```
#   y =        B0 +      B1(x) + B3*(0 or 1)
# EWL = 15.49515 + 0.08482      if "gender" is 0 / Female
# EWL = 12.51547 + 0.08482      if "gender" is 1 / Male
```

b) Display the scatterplot with the superimposed lines

```
plot(EWL ~ age, main="Excess Body Weight Loss (EWL) vs age (by Male and Female)",
     pch=ifelse(gender==1, "M", "W"),
     col=ifelse(gender==1,  3 ,  6))
abline(15.49515, 0.08482, col=6) # Woman
abline(12.51547, 0.08482, col=3) # Man
```



c) Determine the coefficient of determination

```
# According to the above summary in 4.a., the R**2 is 0.07139.
# This says that there is not a strong correlation between the EWL and age.
```

d) Predict the excess body weight (EWL) for 47 years old male

```
predict(model4, data.frame(age=47, gender=1), interval="conf")
```

```
##       fit      lwr      upr
## 1 16.5019 13.40691 19.59689
```

```
# Predicted EWL for a 47-year old Male: 16.5019
```