Exercise 1: Use the data described in problem 12.1. Do the following:

a. Create a table that gives the number of trees that survived and the number that died of each of the nine species.

Solution:
The following r script generates the desired table,
Code:

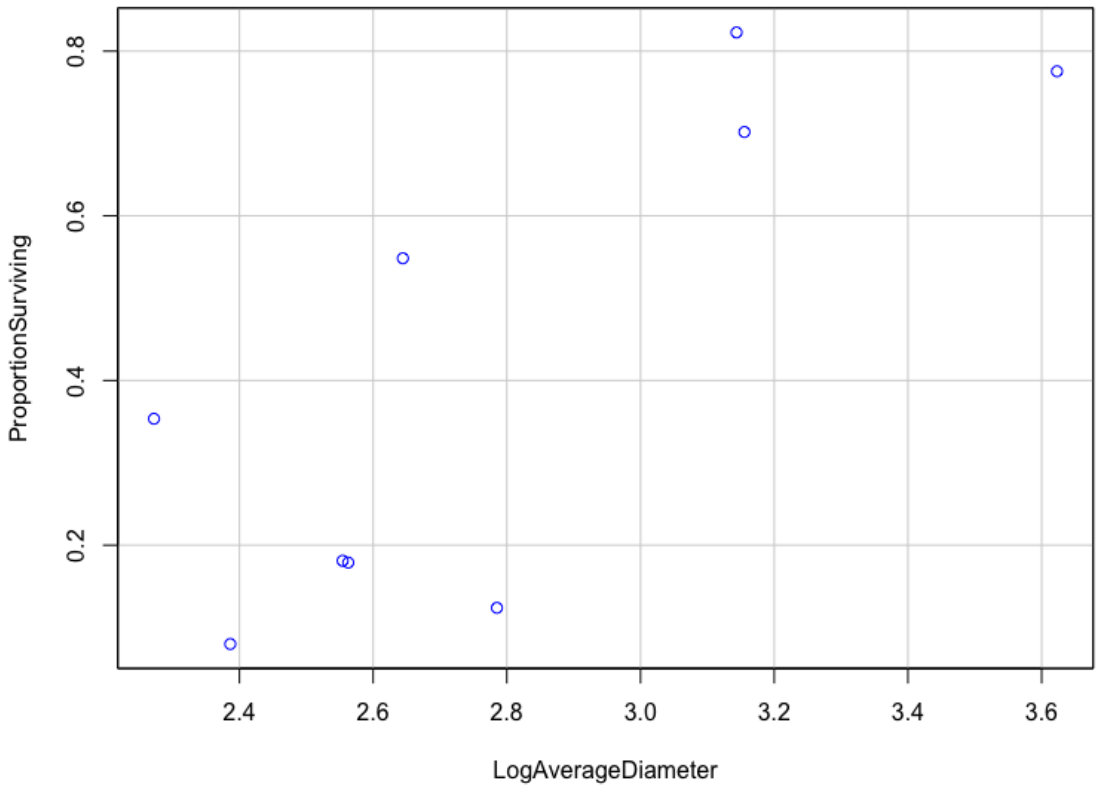
```
df <- Blowdown
ListofFactors <- unique(df$spp)

died = rep(NaN, length(ListofFactors))
survived = rep(NaN, length(ListofFactors))

for (i in 1: length(ListofFactors)){
  died[i] = nrow(subset(df, spp == ListofFactors[i] & y == 1))
  survived[i] = nrow(subset(df, spp == ListofFactors[i] & y == 0))
}

FinalTable <- data.frame(died = died, survived = survived, row.names
```

```
> FinalTable
             died survived
balsam fir      6       69
red pine       90      407
black spruce  233      426
jackpine       44      311
paper birch   413       89
aspen         306      130
red maple      22      101
cedar         532      438
black ash      38       11
```

b. Create a scatter plot tha puts the proportion of deaths in each species on the y-axis and the logarithm, of average diameter on on the x-axis. You can get the surviving proportion in each species using,
aggregate(Blowdown$y, by=list(Blowdown$spp),FUN=mean)$x
and the average diameter using
aggregate(Blowdown$d, by=list(Blowdown$spp),FUN=mean)$x
Comment on whether the sigmoid curve of logistic regression appears to fit the data in your scatter plot.

Solution:
Generating the scatterplot we can see that a sigmoid curve might be able to fit the data, with low proportions near the where the log average diameter is in the high 2s and higher proportions in the low 3s.

Figure 1: Scatterplot of Log average diameter vs survival probability.



c. Fit the logistic regression model to the raw data using $log(d)$ as the regressor. Draw effects plots of the fitted model.

Solution:
Fitting the logistic regression in r and generating the effects plot we get the following,

Figure 2: Plot of Logistic Regression

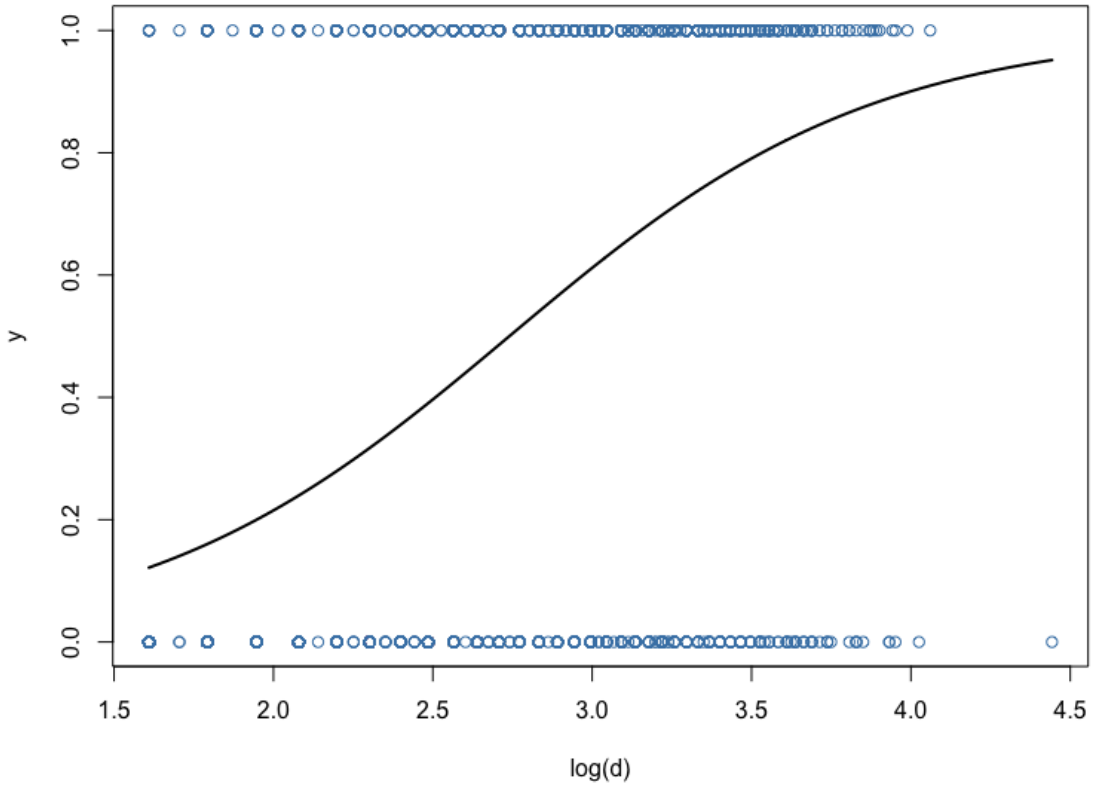
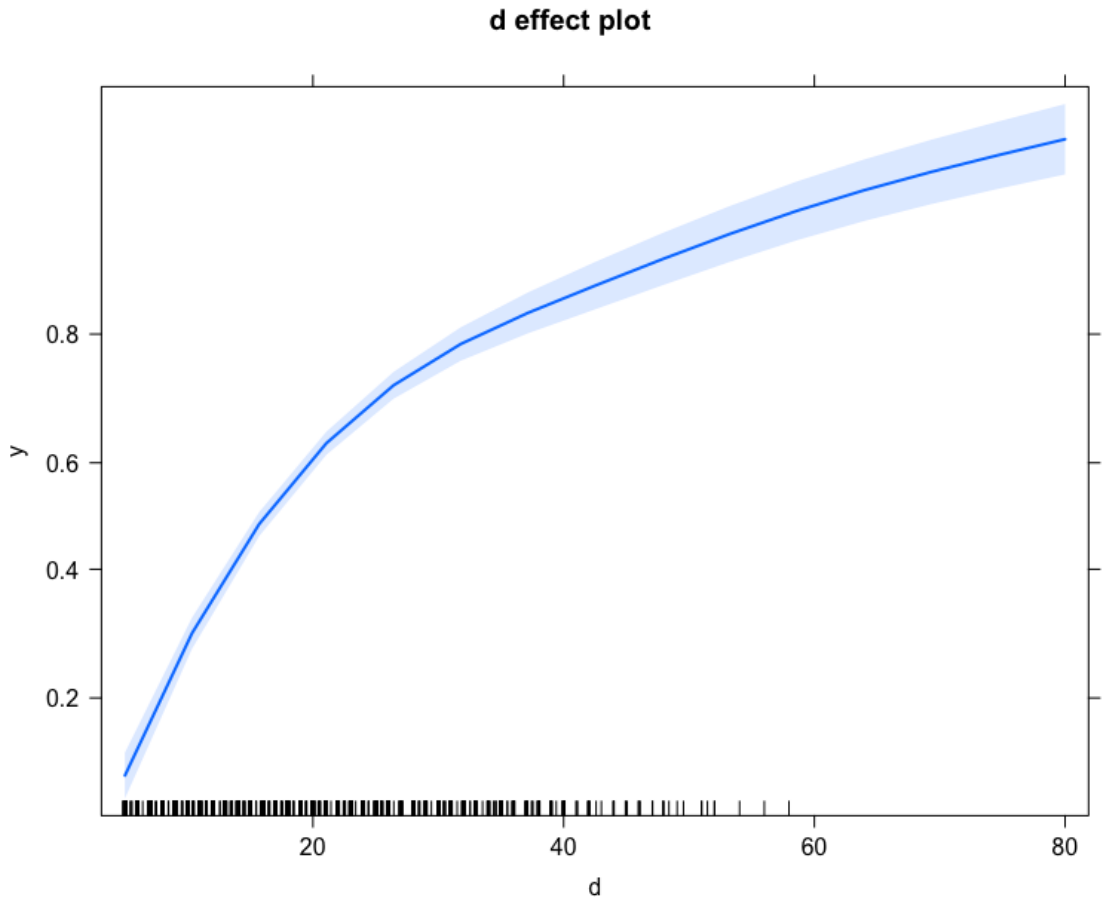


Figure 3: Effects Plot



Code:

```
## Fitting the logisitc regression
> LogisticRegression <- glm(y ~ log(d), data = df, family = binomial)
> summary(LogisticRegression)

Call:
glm(formula = y ~ log(d), family = binomial, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4606  -0.9947  -0.5093   1.0631   2.0527

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.79181    0.20786  -23.05   <2e-16 ***
log(d)       1.74882    0.07678   22.78   <2e-16 ***
---
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5057.9  on 3665  degrees of freedom
Residual deviance: 4417.6  on 3664  degrees of freedom
AIC: 4421.6

Number of Fisher Scoring iterations: 4

## Generating Effects Plot
> plot(allEffects(LogisticRegression))

## Plotting logistic Regression against data.
> plot(y ~ log(d), data=df, col="steelblue")
> xlinspace <- data.frame(d=seq(min(df$d), max(df$d),len=500));
> yValues = predict(LogisticRegression, xlinspace, type="response");
> dff <- data.frame(y = yValues, d = xlinspace)
> lines(y ~ log(d), data=dff, col="steelblue")
```

d. Using the fitted model, give an interpretation of the coefficient for log(d)

Solution:
By looking at the regression summary report we get that as the log of the diameter increases by one unit, the log odds of a tree dying increase by 1.74882.

e. Add $(log(d))^2$ ot the mean funciton form the fitted model to allow for a possible decline in the probabliity of blowdown for the larges trees. Obtain the likelihood ration test for the hypothesis that the quadratic term is 0 and interpret its results.

Solution:
Fitting the second order model and performing the likelihood ratio test we get p-value on the order of 10^{-14} which means that we reject the null hypothesis and conclude that the second order is a significant predictor.
Code:

```
> SecondOrderLogisticRegression <- glm(y ~ log(d) + I(log(d)^2),
                                       data = df,
                                       family = binomial('logit'))

> LogisticRegression <- glm(y ~ log(d),
                            data = df,
                            family = binomial('logit'))

> Anova(SecondOrderLogisticRegression)

Response: y
            LR Chisq Df Pr(>Chisq)
log(d)        99.522  1  < 2.2e-16 ***
I(log(d)^2)   56.036  1  7.115e-14 ***
---
```

Exercise 2: Do problem 12.8. then the problem says 'summarize results', predict the probability of death of a tree with diameter 21cm and local severity measure .5.

12.8.1 For the blowdown example, fit the model

$$y \approx log(d) + s + log(d) : s$$

for spp = paper birch and summarize results.

Solution:

Code:

```
> dfBirch <- subset(df, spp == 'paper birch')
> LogisticRegression <- glm(y ~ log(d) + s + log(d):s,
                            data = dfBirch,
                            family = binomial('logit'))
> predict(LogisticRegression, data.frame( d = 21, s = .5), type="response")
        1
0.8846119
```

12.8.2 Do the same for spp = aspen and summarize results.

Solution:

Code:

```
> dfAspen <- subset(df, spp == 'aspen')
> LogisticRegression <- glm(y ~ log(d) + s + log(d):s,
                            data = dfAspen,
                            family = binomial('logit'))
> predict(LogisticRegression, data.frame( d = 21, s = .5), type="response")
        1
0.7555142
```

Exercise 3: Use the data described in problem 12.9. Do the following:

a. Fit a poisson regression model with sex, citizen, and type as predictors and count as the response. interpret the estimated coefficient for each regression.

Solution:
Fitting the model in r, we get the following summary report.
Code:

```
> PoissonRegression <- glm(count ~ sex + citizen + type,
                           data = df,
                           family = poisson('log'))
> summary(PoissonRegression)

Call:
glm(formula = count ~ sex + citizen + type, family = poisson("log"),
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1097  -1.1612   0.1495   1.1267   5.1718

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.59120    0.08258  43.488  < 2e-16 ***
sexMale      0.73967    0.05655  13.081  < 2e-16 ***
citizenUS   -0.12885    0.05300  -2.431  0.01505 *
typeI(Pu)    0.43504    0.08836   4.924 8.49e-07 ***
typeII       0.29005    0.09102   3.186  0.00144 **
typeIII     -0.21019    0.10289  -2.043  0.04107 *
typeIV       0.51177    0.08706   5.878 4.15e-09 ***
typeVa      -0.87452    0.12690  -6.892 5.52e-12 ***
---

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 521.444  on 23  degrees of freedom
Residual deviance:  99.201  on 16  degrees of freedom
AIC: 252.59

Number of Fisher Scoring iterations: 4
```

Interpreting the coefficients we get that if a subject is male their count changes by a multiplicative factor of $e^{0.73967}$. If the subject is a us citizen their count changes by a multiplicative factor of $e^{-0.12885}$. The type coefficients can be interpreted similarly.

b. Perform a goodness-of-fit test on the model using residual deviance. Interpret the test's result.

Solution:
Computing the p-value for a goodness-of-fit test using residual deviance using r we get a p-value on the order of 10^{-14} so we reject the null hypothesis and conclude that our model does not adequately fit the data.
Code:

```
> gft <- 1-pchisq(99.201,16)
> gft
[1] 4.884981e-14
```