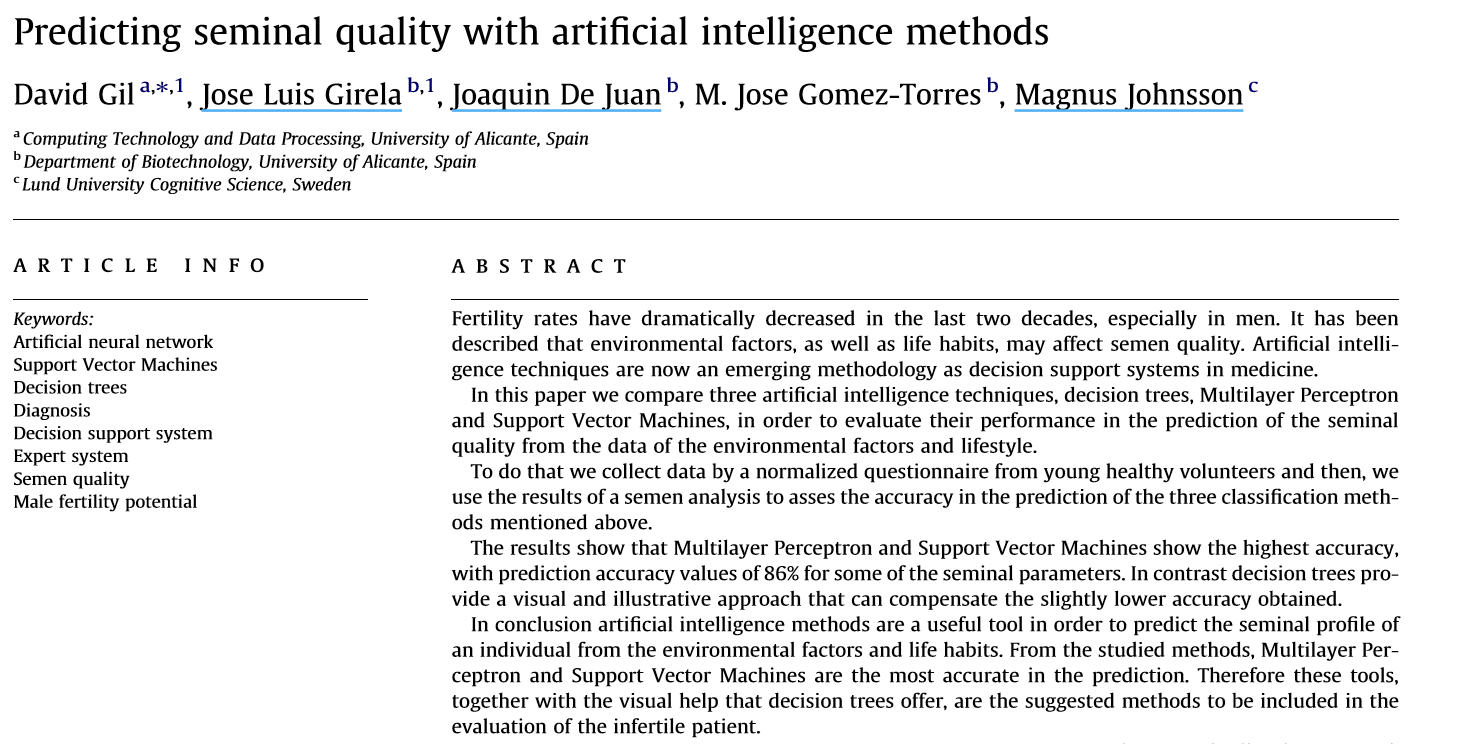
Sarah Brownell & Eleanor Humphreys

Project 2 Report

1. Research Question
   1. Accounting for smoking habits, how does the frequency of alcohol consumption effect the sperm diagnosis?
2. Definition of included variables

|  |  |  |  |
| --- | --- | --- | --- |
| Feature Description | Our Description | Values | Normalized |
| Season |  | 1. Winter, (2) spring, (3) Summer, (4) fall | (-1,-0.33, 0.33, 1) |
| Age |  | 18-36 | (0, 1) |
| Childish Diseases (i.e., chicken pox, measles, mumps, polio) |  | 1. Yes, (2) no | (0, 1) |
| Accident |  | 1. Yes, (2) no | (0, 1) |
| Surgery | Definition of surgery not specified. We will assume surgery is anytime a patient goes fully under anesthesia in order to do a procedure. | 1. Yes, (2) no | (0, 1) |
| High Fevers (in the last year) | Parameters for high fever not specified. | 1. Less than three months age, (2) more than three months age, (3) no | (-1,0,1) |
| Alcohol Consumption |  | Grams per week (0-312) | (0, 1) |
| Smoking Habit |  | 1. never, (2) occasional (3) daily. | (-1,0,1) |
| Sitting Per Day (Number of hours) |  | 1-16 | (0, 1) |
| Diagnosis | To determine if semen was altered or normal the researchers used a semen analysis and compared then the results the reference values put out by the World Health Organization. | (N) Yes, (O) Altered | (N, O) |

1. Description of the study and dataset



Over the past several decades, fertility rates, especially in men have been declining significantly. It is theorized that many distinct factors affect semen quality. In this study data was collected from healthy male participants in the form of a survey. Their semen was collected and then analyzed.

Response Variable= Diagnosis

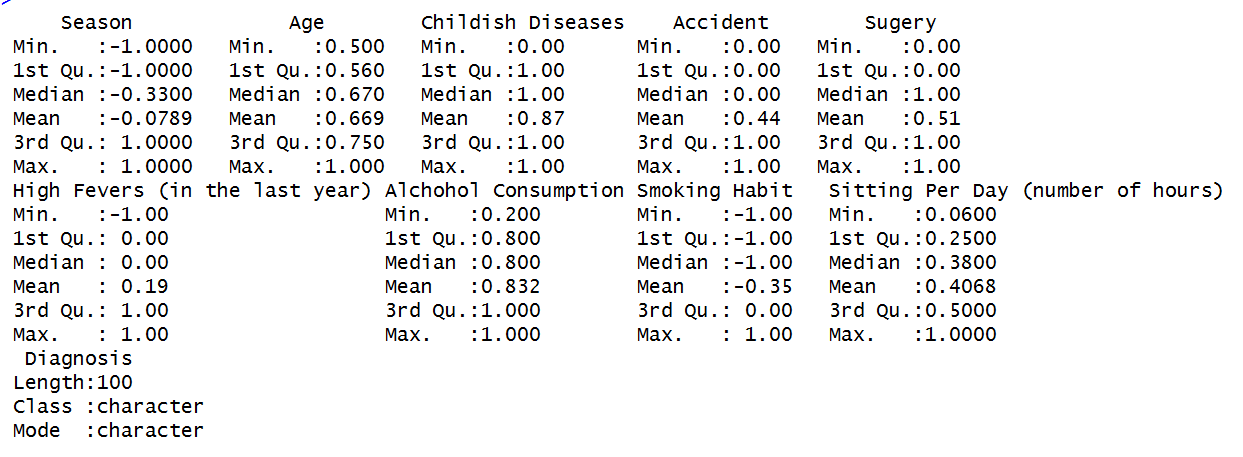
N=Yes (Normal)

0= Altered (Not Normal)

Goal= find a good predictor variable of diagnosis

\*Diagnosis: To determine if semen was altered or normal the researchers used a semen analysis and compared then the results the reference values put out by the World Health Organization.\*

summary(Fertility)



> table(Fertility$Diagnosis)

N O

88 12

4. Summary Information

A. Explanatory Variable= Smoking Habit

X= Smoking Habit(Qual.)

Y= Fertility Diagnosis (0/1)

> View(Fertility)

> df<- rep(0, 100)

> df[Fertility$Diagnosis == "O"]<-1

> table(Fertility$`Smoking Habit`,df)

df

0 1

-1 50 6

0 20 3

1 18 3

Table 1: Percent Abnormal Fertility Diagnosis by Smoking Habits

-1 (Never Smoke) 0 (Occasionally Smoke) 1 (Daily Smoke) All

---------------------------------------------------------------------------------------------------------------------

6% (n=6) 3% (n=3) 3% (n=3) 12% (n=12)

χ2 = 0.21, p = 0.898

> chisq.test(Fertility$`Smoking Habit`, df, correct = FALSE)

Pearson's Chi-squared test

data: Fertility$`Smoking Habit` and df

X-squared = 0.21527, df = 2, p-value = 0.898

The differences in smoking habits were not statistically significant (χ2 = 0.21, p = 0.898).

> t<-table(Fertility$`Smoking Habit`,df)

> p <- 100\*prop.table(t, 2)

> p<- round(p, 1)  
> p

df

0 1

-1 56.8 50.0

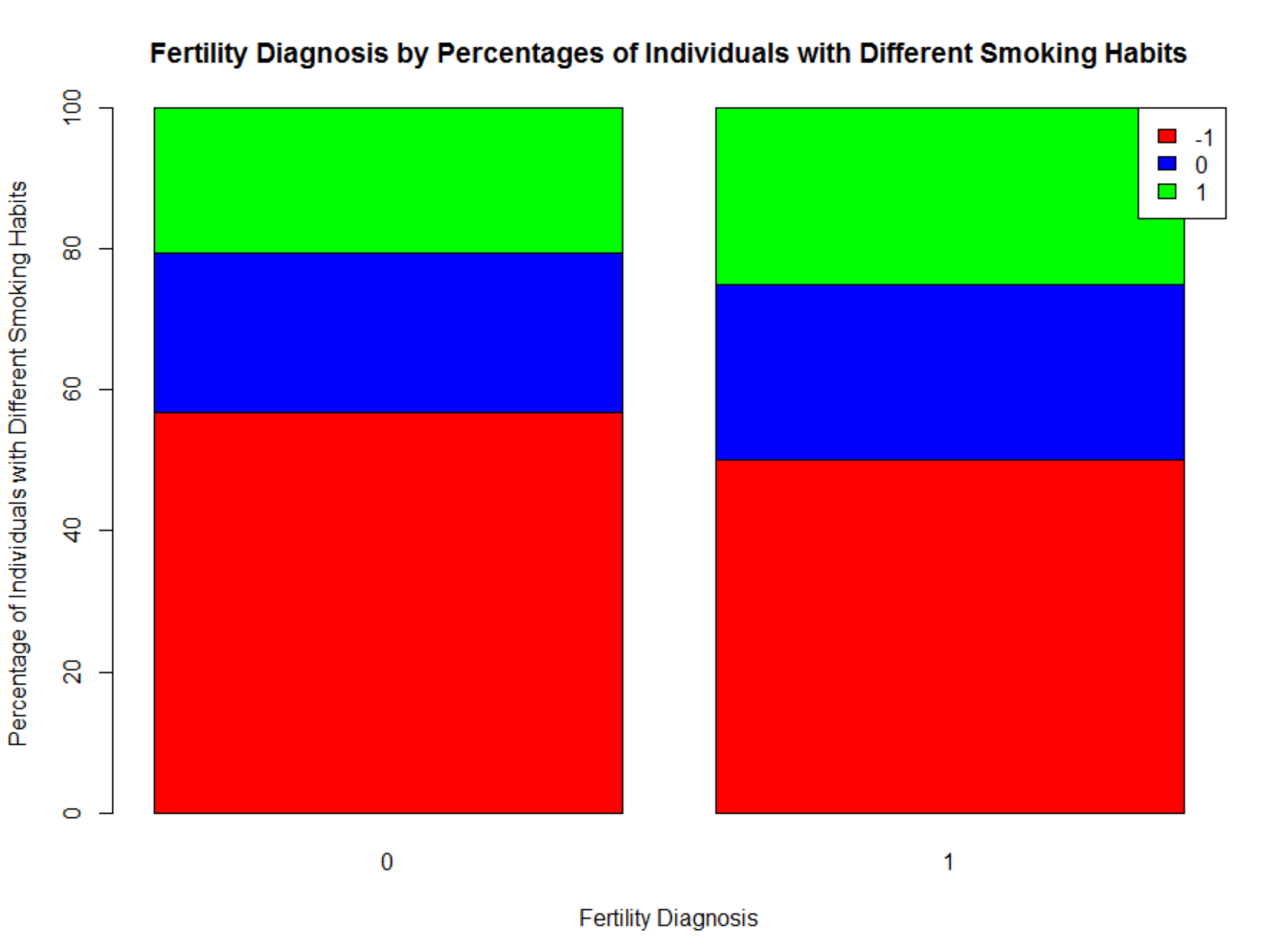
0 22.7 25.0

1 20.5 25.0

> barplot(p, main= "Fertility Diagnosis by Percentages of Individuals with Different Smoking Habits", xlab= "Fertility Diagnosis", ylab="Percentage of Individuals with Different Smoking Habits", col = c("red", "blue", "green"))

> legend("topright", legend = rownames(p),fill = c("red", "blue", "green"))

Figure 1: Fertility Diagnosis by Smoking Habits



The majority of normal and abnormal fertility diagnoses groups are made up of individuals who do not smoke. Of the people with abnormal fertility diagnoses, there are more daily smokers with a difference of 6.8% more daily smokers.

//Interpretation

B. Explanatory Variable= Alcohol Consumption

X= Alcohol Consumption (Qual.)

Y= Fertility Diagnosis (0/1)

> table(Fertility$`Alchohol Consumption`, df)

df

0 1

0.2 1 0

0.4 1 0

0.6 15 4

0.8 33 6

1 38 2

Table 1: Percent Unnormal Fertility Diagnosis by Alcohol Consumption Level

0.2 0.4 0.6 0.8 1 All

--------------------------------------------------------------------------------------------------------------

0% (n=0) 0% (n=0) 4% (n=4) 6% (n=6) 2% (n=2) 12% (n=12)

χ2 = 4.03, p = 0.4025

> chisq.test(Fertility$`Alchohol Consumption`, df, correct = FALSE)

Pearson's Chi-squared test

data: Fertility$`Alchohol Consumption` and df

X-squared = 4.0263, df = 4, p-value = 0.4025

The differences in alcohol consumption were not statistically significant (χ2 = 4.03, p = 0.4025).

t<-table(Fertility$`Alchohol Consumption`,df)

p <- 100\*prop.table(t, 2)

p<- round(p, 1)

> p

df

0 1

0.2 1.1 0.0

0.4 1.1 0.0

0.6 17.0 33.3

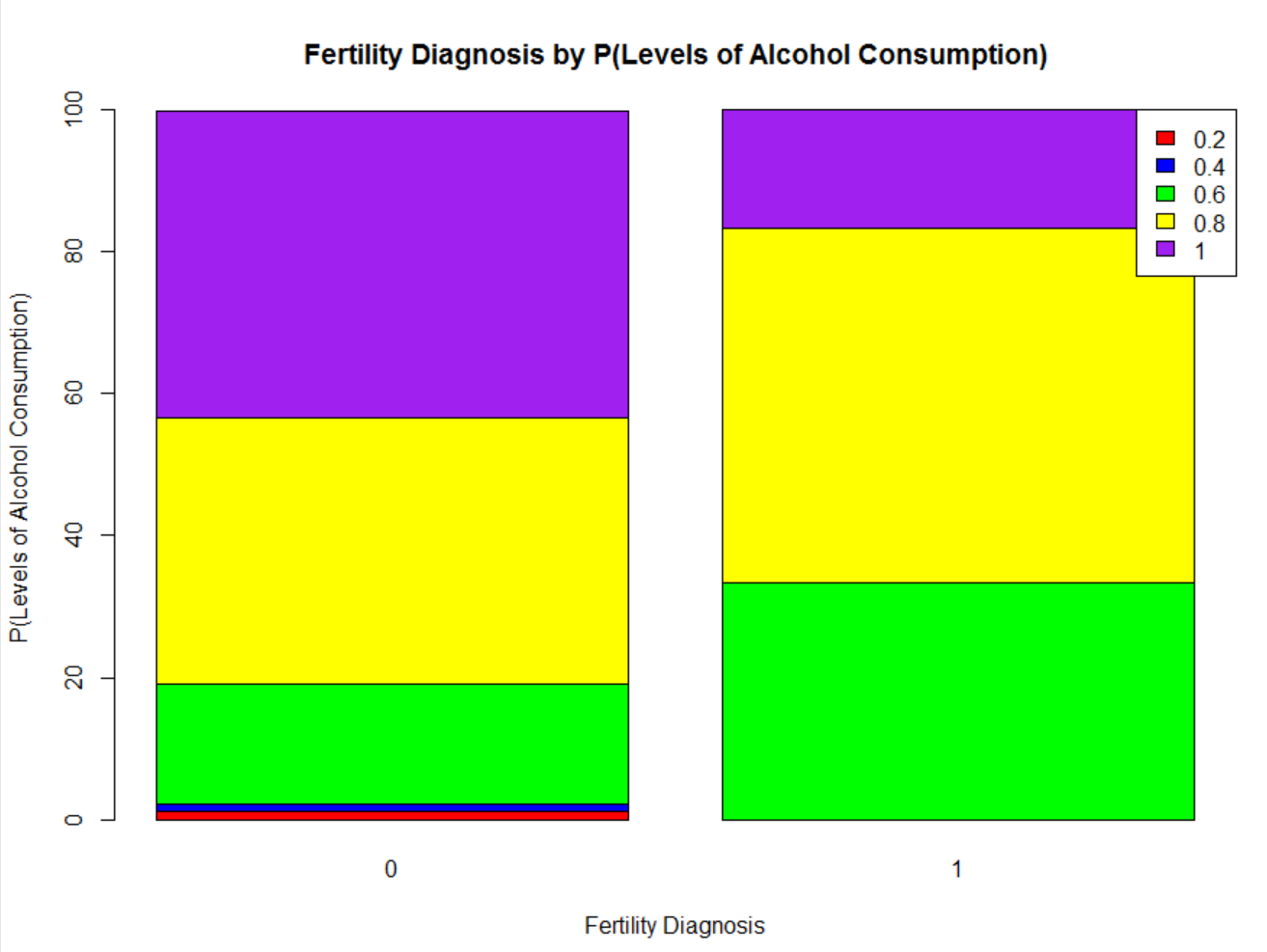
0.8 37.5 50.0

1 43.2 16.7

> barplot(p, main= "Fertility Diagnosis by P(Levels of Alcohol Consumption)", xlab= "Fertility Diagnosis", ylab="P(Levels of Alcohol Consumption)", col = c("red", "blue", "green", "yellow", 'purple'))

> legend("topright", legend = rownames(p),fill = c("red", "blue", "green", "yellow", "purple"))

Figure 2: Alcohol Consumption by Fertility Diagnosis



There is a significant difference among the abnormal and normal fertility diagnoses groups in regards to alcohol consumption. The group of individuals with the highest level of alcohol consumption makes up a larger percentage of the individuals with normal fertility diagnoses than abnormal diagnoses with a difference of 26.5% between the groups.

//Interpretation

5. Models

A. X=Alcohol Consumption, Y=Fertility Diagnosis

> df<- rep(0, 100)

> df[Fertility$Diagnosis == "O"]<-1

> df

> model<- glm(df~Fertility$`Alcohol Consumption`, binomial)

> model

Call: glm(formula = df ~ Fertility$`Alcohol Consumption`, family = binomial)

Coefficients:

(Intercept) Fertility$`Alcohol Consumption`

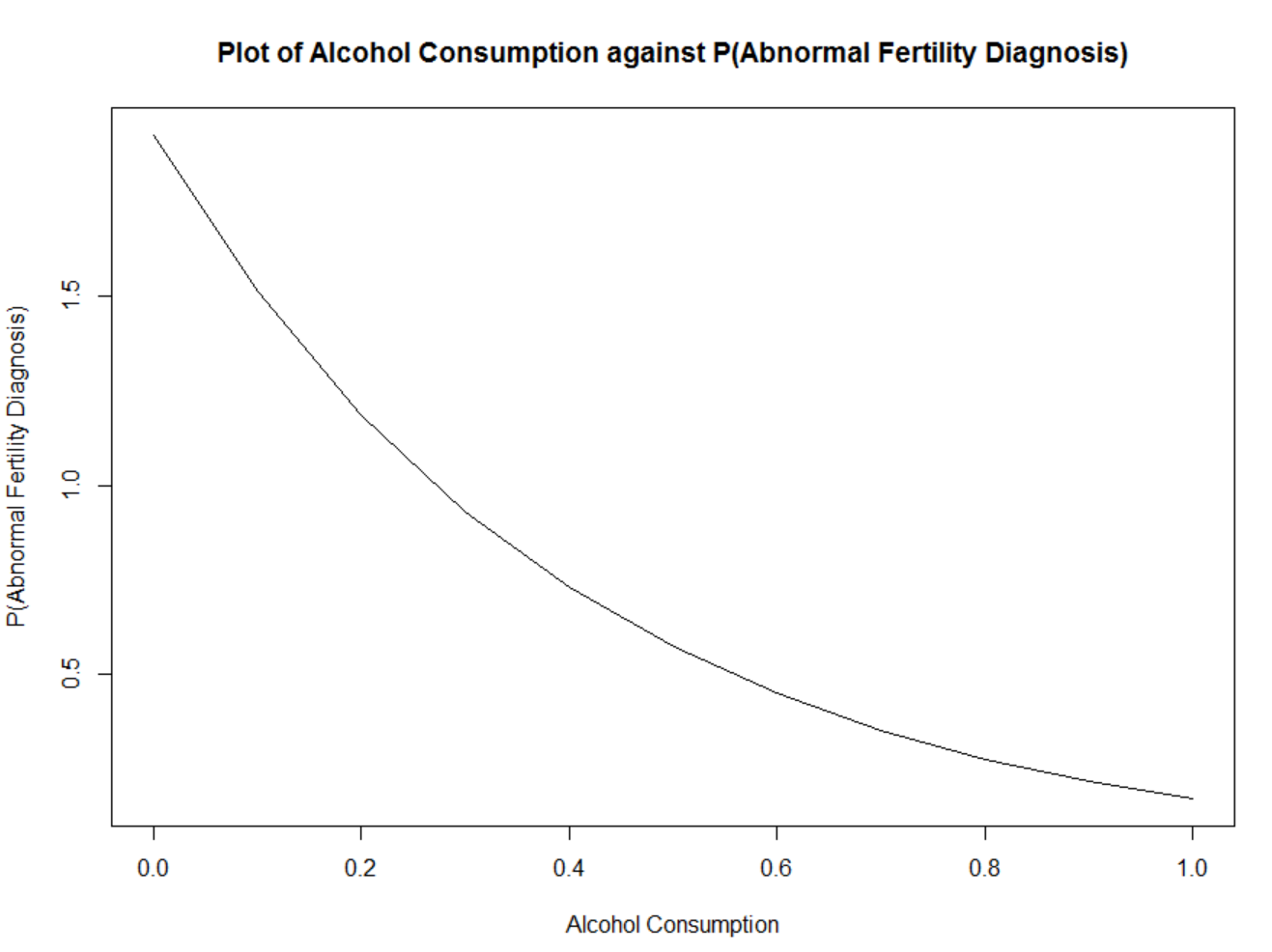
-0.03708 -2.42680

P(Predicted Y=1) =

> x<- seq(0,1,.1)

> y<- exp(-.0371-2.43\*x)/1+exp(-.0371-2.43\*x)

> plot(y~x, type= "l",main= "Plot of Alcohol Consumption against P(Abnormal Fertility Diagnosis)", xlab="Alcohol Consumption", ylab="P(Abnormal Fertility Diagnosis)")



The plot of Alcohol Consumption against P(Y=1) indicates the negative quadratic relationship of fertility diagnosis as the alcohol consumption increases. This means individuals in the study had a larger probability for a not normal fertility diagnosis as average alcohol consumption increased on a normalized scale.

The higher one’s alcohol consumption, the lower the predicted probability of having a not normal fertility diagnosis is.

B. X=Smoking Habit, Y=Diagnosis

> model2<- glm(df~Fertility$`Smoking Habit`, binomial)

> model2

Call: glm(formula = df ~ Fertility$`Smoking Habit`, family = binomial)

Coefficients:

(Intercept) Fertility$`Smoking Habit`

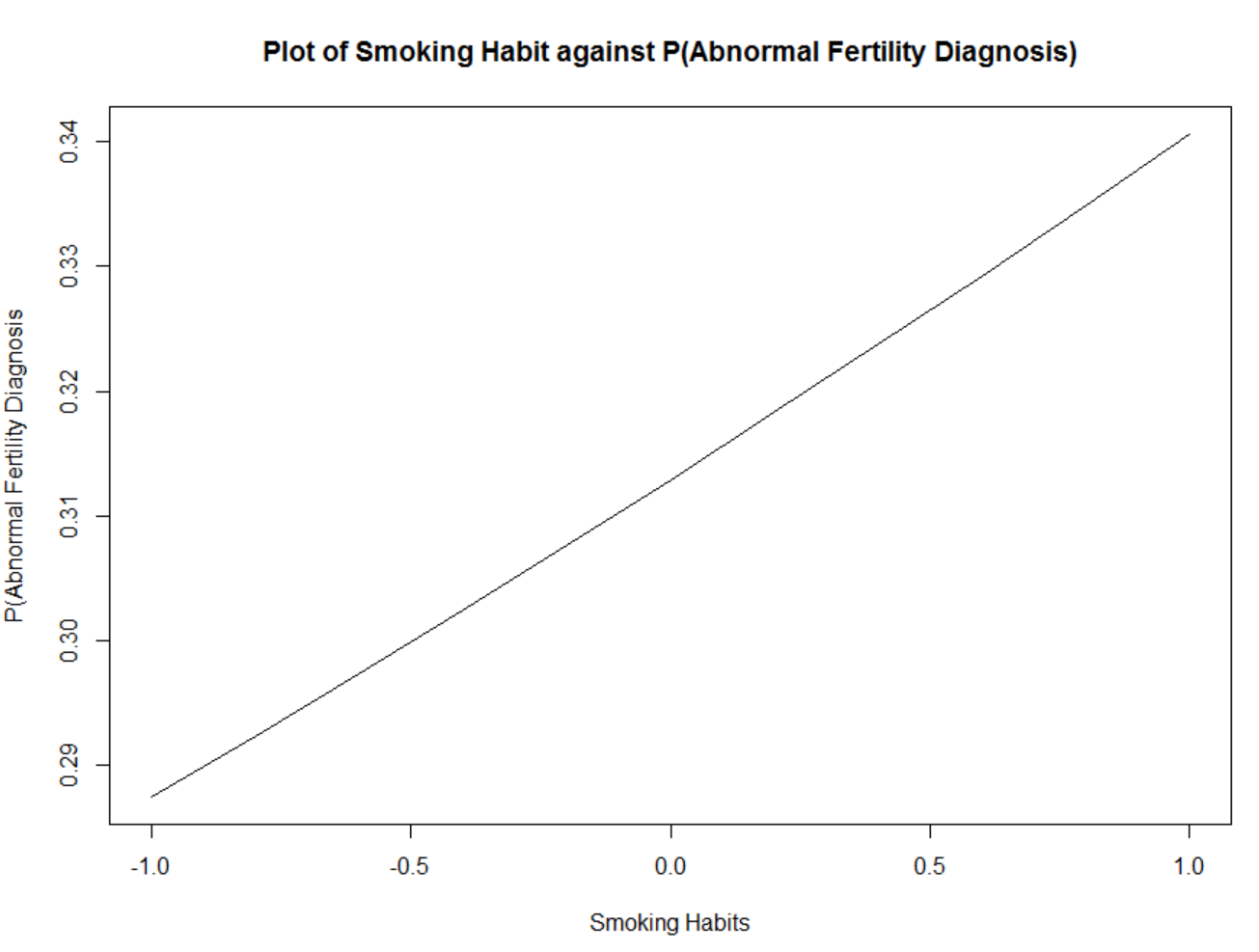
-1.9402 0.1698

P(Predicted Y=1) =

> x1<- seq(-1,1,.2)

> y1<-exp(-1.94+.17\*x)/1+exp(-1.94+.17\*x)

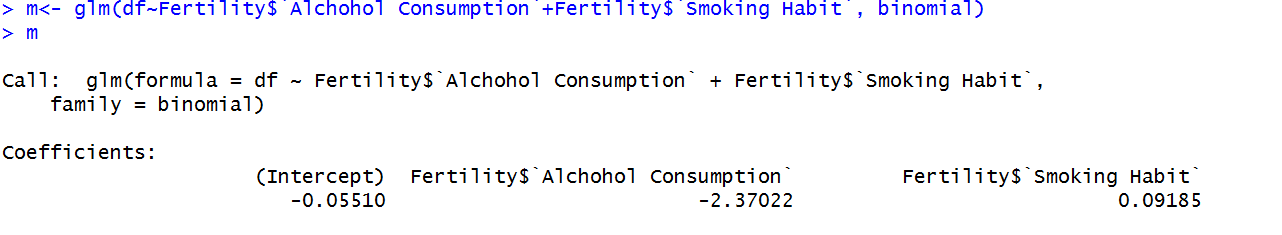
> plot(y1~x1,type="l", main= "Plot of Smoking Habit against P(Abnormal Fertility Diagnosis)", xlab="Smoking Habits",ylab = "P(Abnormal Fertility Diagnosis")



The plot of smoking habit against P(Y=1) indicates a positive linear trend of fertility diagnosis as smoking habit increased. This means that as people smoked more their likelihood of not normal fertility diagnosis increased. However, it increased in a small amount of 0.17 probability of a not normal sperm diagnosis for every 0.5 increase in smoking habit.

The higher one’s smoking habits, the higher the predicted probability of having a not normal fertility diagnosis is.

C. X= Alcohol Consumption, X=Smoking Habit, Y=Diagnosis



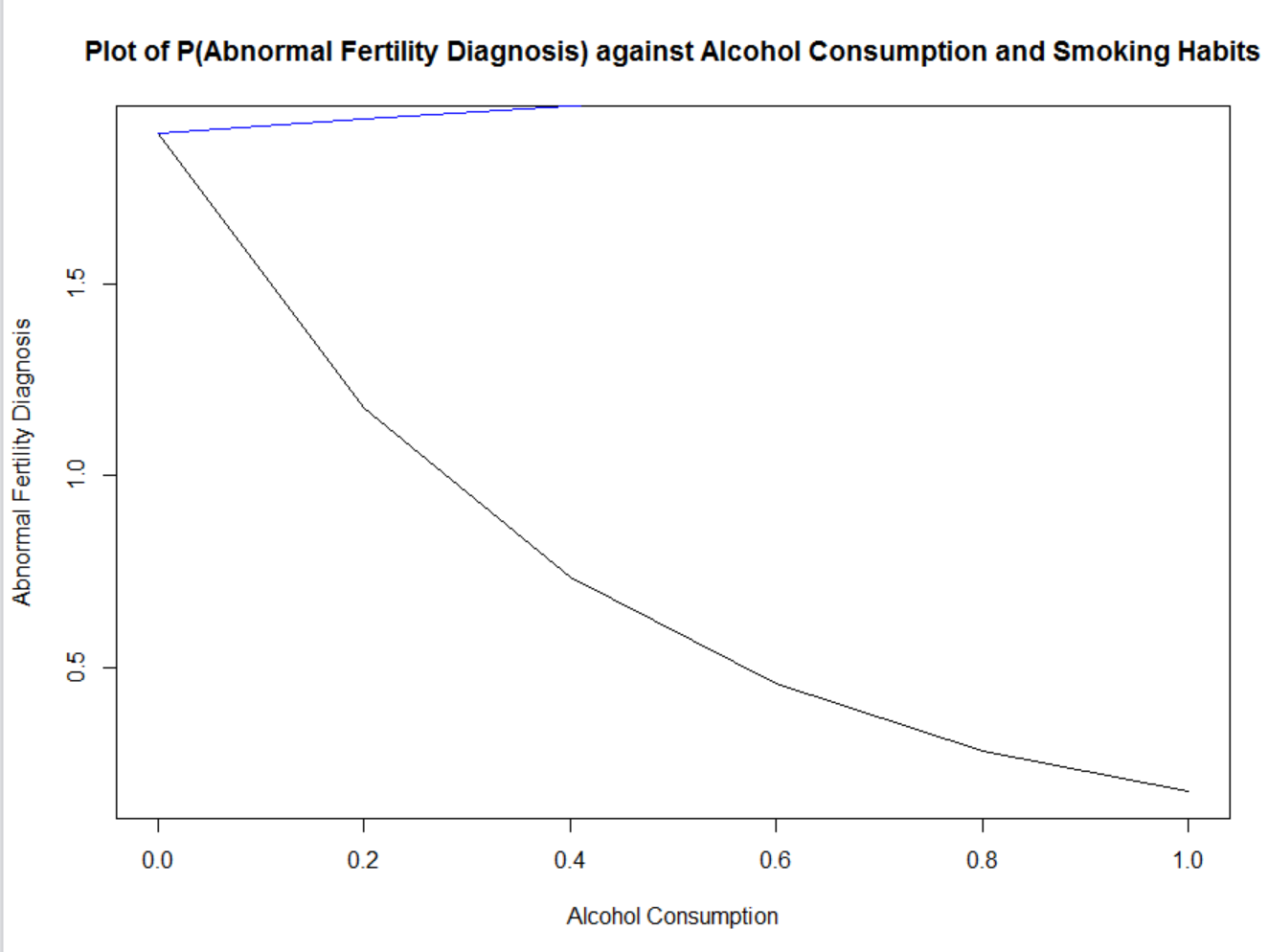
> x1<- seq(0,1,.2)

> y1<-exp(-0.055-2.37\*x1)/1+exp(-0.055-2.37\*x1)

> y2<- exp(-0.055+0.09\*x1)/1+exp(-0.055+0.09\*x1)

> plot(y1~x1, type="l", main= "Plot of P(Abnormal Fertility Diagnosis) against Alcohol Consumption and Smoking Habits", xlab= "Alcohol Consumption", ylab = "Abnormal Fertility Diagnosis")

> lines(x1,y2, col="blue")



6. Conclusions

Limitations

1. Lack of definition of parameters survey type questions

Ex. High fever with no specific number

2. Hard to understand standardized scale of grams per alcohol per week

7. Appendix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Fertility Diagnosis |  |  |
|  |  | 0 (Normal) | 1 (Not Normal) | Total |
| Smoking Habits | -1 (Never Smoke) | 50% (n=50) | 6% (n=6) | 56% (n=56) |
|  | 0 (Occasionally Smoke) | 20% (n=20) | 3% (n=3) | 23% (n=23) |
|  | 1 (Daily Smoke) | 18% (n=18) | 3% (n=3) | 31% (n=31) |
|  | Total | 88% (n=88) | 12% (n=12) | 100% (n=100) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Fertility Diagnosis |  |  |
|  |  | 0 (Normal) | 1 (Not Normal) | Total |
| Alcohol | 0.2 | 1% (n=1) | 0% (n=0) | 1% (n=1) |
| Consumption | 0.4 | 1% (n=1) | 0% (n=0) | 1% (n=1) |
|  | 0.6 | 15% (n=15) | 4% (n=4) | 19% (n=19) |
|  | 0.8 | 33% (n=33) | 6% (n=6) | 39% (n=39) |
|  | 1 | 38% (n=38) | 2% (n=2) | 40% (n=40) |
|  | Total | 88% (n=88) | 12% (n=12) | 100% (n=100) |