MATH 536 HW 5

Gustavo Esparza

10/30/2019

1 Donner Party

A Fit a logistic regression model for survival as a function of age. Write out the fitted model.

i Interpret the coefficient of the estimated effect of age. Do the odds of survival increase or decrease with age?

Here are the coefficient results from our GLM:

```
##                Estimate Std. Error   z value   Pr(>|z|)
## (Intercept)  1.81851831 0.99937233  1.819660 0.06881073
## age         -0.06647028 0.03222003 -2.063011 0.03911155
```

Using the logit coefficients, we have the following logistic model:

$$logit(\pi) = 1.82 - .066(AGE)$$

The coefficient of the estimated effect of age is -0.06647. Taking the exponent of this value gives us $e^{-.06647} = 0.935691$. This is telling us that the odds for survival is 0.935691 times the odds for survival when considering a one unit increase in age. As the coefficient for age is negative, the odds of survival decreases as age increases.

ii Determine the probability of survival for a newborn (age = 0).

$$P(survival|age = 0) = \frac{exp(1.81852 - 0.06647. \times 0)}{1 + exp(1.81852 - 0.06647 \times 0)} = \frac{exp(1.81852)}{1 + exp(1.81852)}$$

By using an age of zero for our logistic regression model, the probability of survival for a newborn is 0.8603882.

iii Compute a 95% confidence interval for the odds ratio for age. What is the formal interpretation of this confidence interval?

The odds for suvival multiply by $e^{-.06647}$ for every 1 unit increase in age, thus giving the result:

```
##
## Change in Odds From a One Unit Increase in Age =  0.9356907
```

Considering a normal assumption for our regression coefficients, we have the following confidence interval for the logistic regression coeffeient for age,using the provided standard error:

$$-.06647 \pm 1.96 \times 0.03222$$

Taking the exponent of our endpoints, we have the following:

```
##
## Confidence Interval = [ 0.8784278  ,  0.9966865 ]
```

The odds for survival is estimated to be between 88% and 99% the odds when the age is increased by one unit.

iv Conduct a test of lack of fit for this model based on deviance. Be sure to specify your null and alternative hypotheses, test statistics, distribution of the test statistic under the null, p-value, and your conclusions.

Hypothesis:
H_0 : Intercept only model (Survival is independent of Age) VS H_a : Model including age (Survival is dependent on Age)

Test statistic:

ΔG^2 = Deviance for model not including age (null deviance) - deviance for model including age (residual deviance) = 61.8265419 - 56.2907216 = 5.5358203 .

Distribution under null: χ^2 with df = 2 parameters - 1 parameter = 1

p-value: 0.018631

Thus, we proceed to reject the null hypothesis and conclude that Age is significant to detrmining Survival Rate.

B Consider an additive model with the variables age and sex. Fit a logistic regression model for survival as a function of age and sex. Write out the fitted model.

i Interpret the coefficient of the estimated effect of gender.

Here are the coefficient results for GLM considering Age and Gender as predictors:

```
##                Estimate Std. Error   z value   Pr(>|z|)
## (Intercept)  3.23041381  1.3868611  2.329299 0.01984325
## age         -0.07820407  0.0372845 -2.097496 0.03594972
## gender      -1.59729350  0.7554690 -2.114307 0.03448904
```

Using the logit coefficients, we have the following logistic model:

$$logit[\pi] = 3.23 - 0.078 \times \textbf{age} - 1.597 \times \textbf{gender}$$

We can observe that when considering a gender of male as opposed to female, the change in odds of survival by a factor of $e^{-1.59729} = 0.2024444$, when age is held constant. That is, the odds of survival for males is about .2 times the odds of survival for females, considering age is held constant.

ii Determine the probability of survival for a 24 year old male.

P(survival|age = 24, gender = Male) = $\frac{exp(3.231-0.08\times 24-1.59\times 1)}{1+exp(3.231-0.08\times 24-1.59\times 1)}$

The probability of survival for a 24 year old male is 0.4393557.

iii Determine the probability of survival for a 24 year old female.

P(survival|age = 24, gender = Female) = $\frac{exp(3.231-0.08\times24)}{1+exp(3.231-0.08\times24)}$

The probability of survival for a 24 year old female is 0.7947039.

iv Conduct a test to determine if the variable sex should be included in the model.

Hypothesis: H_0 : Model only including age VS H_a : Model including age and sex

Test statistic: Δg^2 = Deviance for model not including sex - deviance for model including sex = 5.034437.

Distribution under null: χ^2 with df = (3 parameters) - (2 paramters) = 1

p-value: 0.0248482

Thus, we proceed to reject the null hypothesis and conclude that Sex is significant to detrmining Survival Rate, in additon to age.

v What are the odds of survival for a 50 year old female to a 20 year old female?

The odds of survival for a 50 year old female to a 20 year old female is defined as $e^{0.07820\times(50-20)} = e^{0.07820\times30}$, giving the following

```
## odds of survival for a 50 year old female to a 20 year old female =  0.09573971
```

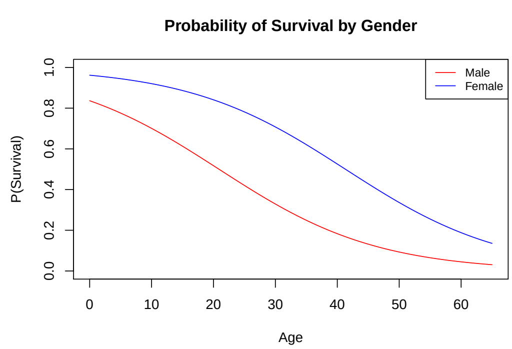
vi Write two models: one model for males and one model for females. Plot the model for males and the model for females on the same graph. Based on these models, are women more likely to survive than men?}

Here are the two probability models for each respective gender:

Model for males: $\pi = \frac{exp\left((3.23041-1.59729)-0.07820\times\textbf{age}\right)}{1+exp\left((3.23041-1.59729)-0.07820\times\textbf{age}\right)}$

Model for females: $\pi = \frac{exp\left(3.23041-0.07820\times\textbf{age}\right)}{1+exp\left(3.23041-0.07820\times\textbf{age}\right)}$

Here are the plots for our two models:



Based on this plot, we can see that females have a consistently higher probability of survival when compared to men.

C Finally, test for the interaction between age and gender by fitting a logistic regression model for survival as a function of age and sex. Write out the fitted model. Be sure to specify your null and alternative hypotheses, test statistics, distribution of the test statistic under the null, p-value, and your conclusions. Plot the estimated model. What are your conclusions?

```
##                Estimate Std. Error   z value   Pr(>|z|)
## (Intercept)  7.2463848 3.20516725  2.260845 0.02376889
## age         -0.1940741 0.08741539 -2.220137 0.02640948
## gender      -6.9280495 3.39887166 -2.038338 0.04151614
## age:gender   0.1615969 0.09426191  1.714339 0.08646650
```

Using the logit coefficients, we have the following logistic model:

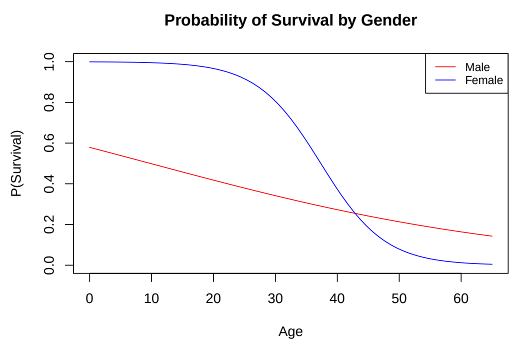
$$logit[\pi] = 7.2463848 - 0.1940741 \times \textbf{age} - 6.9280495 \times \textbf{gender} + 0.1615969 \times \textbf{age : gender}$$

When considering gender, we have the following probability models:

Model for males: $\pi = \frac{exp\left(.318 - .032 \times \textbf{age}\right)}{1 + exp\left(.318 - 0.032 \times \textbf{age}\right)}$

Model for females: $\pi = \frac{exp\left(7.24 - .194 \times \textbf{age}\right)}{1 + exp\left(7.24 - .194 \times \textbf{age}\right)}$

Using these models, here are the interaction model plots given age:



Here, we can see that the probability of survival for females is initially higher than the probability of survival for males. As age increases, we find that eventually males have a higher probability of survival when compared to females of the same age.

Hypothesis: H_0 : Model only including age and gender VS H_a : Model including age, gender, *and* age:gender interaction.

Test statistic: Δg^2 = Deviance for model not including age/gender interaction - deviance for model including age/gender interaction = 3.9099182.

Distribution under null: χ^2 with df = (4 parameters) - (3 paramters) = 1

P-value: 0.0480019

Thus, we proceed to reject the null hypothesis and conclude that the interaction term between age and gender is significant to the model.

2 Book Problems

4.3

In the first nine decades of the twentieth century in baseball's National League, the percentage of times the starting pitcher pitched a complete game were: 72.7 (1900-1909), 63.4, 50.0, 44.3, 41.6, 32.8, 27.2, 22.5, 13.3 (1980-1989) (Source: George Will, Newsweek, April 10, 1989).

A

Treating the number of games as the same in each decade, the linear probability model has ML fit $\hat{\pi}$ = 0.7578 - 0.0694 x , where x = decade (x = 1, 2, . . . , 9). Interpret -0.0694.

-.0694 can be interpreted as the estimates decrease in probability of pitching a complete game as the decade increases by one unit.

B

Substituting x = 12, predict the percentage of complete games for 2010 - 2019. Is this prediction plausible? Why?

The percentage of complete games for 2010-2019 is -0.075. As this prediction value is not between 0 and 1, thus it is not possible to have a negative probability.

C

The logistic regression ML fit is $\hat{pi}$ = exp(1.148 - 0.315 x)/[1 + exp(1.148 - 0.315 x)]. Obtain $\hat{\pi}$ for x = 12. Is this more plausible than the prediction in (b)?

$\hat{\pi}$ for x =12, the predicted percentage of complete games for 2010-2019 is 0.07. This prediction indicates that the probability of a complete games is about 7 %, which is a much more plausible prediction.

4.4

Consider the snoring and heart disease data of Table 3.1 in Section 3.2.2., With scores 0,2,4,5 for snoring levels, the logistic regression ML fit is logit($\hat{pi}$) = -3.866 + 0.397x.

A

Interpret the sign of the estimated effect of x.

As the coefficient associated with x is .397, we estimate that the logit probability increases by .397 as the snoring level increases by one unit. We estimate an increase in probability since this coefficient is positive.

B

Estimate the probabilities of heart disease at snoring levels 0 and 5.

Using the logit function, $\pi = exp(-3.866 + .397x)/1 + exp(-3.866 + .397x)$,the estimated probability of heart disease at snoring level 0 is 0.0205124 and the estimated probability of heart disease at snoring level 5 is 0.1322741.

C

Describe the estimated effect of snoring on the odds of heart disease.

The odds for heart disease are denoted by

$$\frac{\pi(x)}{1-\pi(x)} = exp(-3.866 + .397 \cdot x) = exp(-3.866) \times exp(.397)^x$$

We can see that for every unit increase in x, the odds are multiplied by the quantity $exp(.397) = 1.487$. Thus, for each unit increase in snoring level, the odds of heart disease increase by 1.487.

4.10

An international poll quoted in an Associated Press story (December 14, 2004) reported low approval ratings for President George W. Bush among traditional allies of the United States, such as 32% in Canada, 30% in Britain, 19% in Spain, and 17% in Germany. Let Y indicate approval of Bush's performance (1 = yes,0 = no),π = P(Y = 1),c_1 = 1 for Canada and 0 otherwise,c_2 = 1 for Britain and 0 otherwise, and c_3 = 1 for Spain and 0 otherwise.

A

Explain why these results suggest that for the identity link function, $\hat{\pi} = 0.17 + 0.15c_1 + 0.13c_2 + 0.02c_3$**.**

The Identity Link is used for simple linear models and therefore is modeled by $g(\mu) = \mu$. Since no c_i value was assigned to Germany as a response., we can see its respective probability of .17 indicated by the intercept of our link function. Thus, each coefficient for the defined c_i variables will simply be the difference between the intercept and the probabilities provided. Specifically,

$$\begin{aligned}
&\text{Canada: } \beta_1 = .32 - .17 = .15\\
&\text{Britian: } \beta_2 = .30 - .17 = .13\\
&\text{Spain: } \beta_3 .19 - .17 = .02
\end{aligned}$$

Which results in the identity link provided.

B

Show that the prediction equation for the logit link function is logit($\hat{\pi}$) = $-1.59+0.83c_1+0.74c_2+0.14c_3$**.**

The logit link is defined as

$$logit(\hat{\pi}) = ln\Big(\frac{\hat{\pi}}{1-\hat{\pi}}\Big)$$

This provides the following values for our logit link function:

$$\text{Germany: } ln\Big(\frac{.17}{1-.17}\Big) = 1.59$$

Once again, this provides the intercept for our logit link function.

Now for the probabilities defined by c_i:

$$\text{Canada: } ln\left(\frac{.32}{1-.32}\right) = -.75$$

Adding the logit for Canada to our Germany intercept is $(-.75 + 1.59 = .83)$, thus providing our given coefficient.

Similarly, Britain and Spain provides the following:

$$\text{Britian: } ln\left(\frac{.30}{1-.30}\right) = -.84$$

Adding the logit for Britain to our Germany intercept is $(-.84 + 1.59 = .74)$, thus providing our given coefficient.

$$\text{Spain: } ln\left(\frac{.19}{1-.19}\right) = -1.45$$

Adding the logit for Britain to our Germany intercept is $(-1.45 + 1.59 = .14)$, thus providing our given coefficient.

Combining our calculated coefficients provide the desired prediction equation.

4.19

A sample of subjects were asked their opinion about current laws legalizing abortion (support, oppose). For the explanatory variables gender (female, male), religious affiliation (Protestant, Catholic, Jewish), and political party affiliation (Democrat, Republican, Independent), the model for the probability π of supporting legalized abortion

$$logit(\pi) = \alpha + \beta_h^G + \beta_i^R + \beta_j^P$$

has reported parameter estimates (setting the parameter for the last category of a variable equal to 0.0)

$\hat{\alpha} = -0.11, \hat{\beta}_1^G = 0.16, \hat{\beta}_2^G = 0.0, \hat{\beta}_1^R = -0.57, \hat{\beta}_2^R = -0.66, \hat{\beta}_3^R = 0.0, \hat{\beta}_1^P = 0.84, \hat{\beta}_2^P = -1.67, \hat{\beta}_3^P = 0.0.$

A

Interpret how the odds of supporting legalized abortion depend on gender.

The difference for our estimate logit values for gender is .16 - 0 = .16. As .16 is the parameter for the Female gender, we have the odds impact for gender being $exp(.16) = 1.17$. Thus, the odds of a female supporting legalized abortion is about 1.17 times the odds of a male supporting legalized abortion.

B

Find the estimated probability of supporting legalized abortion for (i) male Catholic Republicans and (ii) female Jewish Democrats.

The probability of supporting legalized abortion given our parameters is defined as

$$\pi = \frac{exp(\alpha + \beta_h^G + \beta_i^R + \beta_j^P)}{1 + exp(\alpha + \beta_h^G + \beta_i^R + \beta_j^P)}$$

For male catholic republicans, we have

$$\pi = \frac{exp(-.11 + 0 - .66 - 1.67)}{1 + exp(-.11 + 0 - .66 - 1.67)} = .08$$

For female Jewish Democrats, we have:

$$\pi = \frac{exp(-.11 + .16 + 0.84)}{1 + exp(-.11 + .16 + 0.84)} = .71$$

C

If we defined parameters such that the first category of a variable has value 0, then what would $\hat{\beta}_2^G$ equal? Show then how to obtain the odds ratio that describes the conditional effect of gender.

By defining the parameters as the first category having value zero, then $\hat{\beta}_2^G$ would simply equal $\hat{\beta}_1^G$ -.16.

Since the odds ratio is still computed using the difference in logit coefficients, we have 0 - (-.16), still providing .16. Thus, the odds ratio for the effect of gender remains as $exp(.16) = 1.17$.

D

If we defined parameters such that they sum to 0 across the categories of a variable, then what would $\hat{\beta}_1^G$ and $\hat{\beta}_2^G$ equal? Show then how to obtain the odds ratio that describes the conditional effect of gender.

By setting the parameters to sum to zero, but still maintaining a difference of .16 will provide the following parameters:

$$\hat{\beta}_1^G = .08$$
$$\hat{\beta}_2^G = -.08$$

As the difference still remains as .08 - (-.08) = .16, we observe the same odds ratio to describe the conditional effect of gender.

4.37

For data from Florida on Y = whether someone convicted of multiple murders receives the death penalty (1 = yes, 0 = no), the prediction equation is logit($\hat{\pi}$) = -2.06 + .87d - 2.40v, where d and v are defendant's race and victims' race (1 = black, 0 = white). The following are true-false questions based on the prediction equation.

A

The estimated probability of the death penalty is lowest when the defendant is white and victims are black.

True

The coeffecient associated with defendant is .87 and the coeffecient associated with victim is -2.4. For this logit to be minimized, the defendant coefficient would need to be inactive and the victim coeffecient would need to be active. This gives the case of a white defendant and a black victim.

B

Controlling for victims' race, the estimated odds of the death penalty for white defendants equal 0.87 times the estimated odds for black defendants. If we instead let d = 1 for white defendants and 0 for black defendants, the estimated coefficient of d would be 1/0.87 = 1.15 instead of 0.87.

False

The estimated odds is determined by the exponent of the provided logit coeffecients. Furthermore, by changing the defendant race coding, the estimated odds would be defined as $1/e^{.87}$ and then the coeffecient of d would be -.87.

C

The lack of an interaction term means that the estimated odds ratio between the death penalty outcome and defendant's race is the same for each category of victims' race.

True

When considering the two cases of victim rate, the intercept term changes but the coefficient for defendant race remains as $-.87$. As this slope term determines the odds ratio, we can see that it remains the same for both victim races. If an interaction term was introduced, then the slope term would change based on the race of the victim.

D

The intercept term -2.06 is the estimated probability of the death penalty when the defendant and victims were white (i.e., d = v = 0).

False

The probability needs to be between 0 and 1, so a value of -2.06 is impossible. -2.06 is defined as the *logit* value that can be used to estimate the probability.

E

If there were 500 cases with white victims and defendants, then the model fitted count (i.e., estimated expected frequency) for the number who receive the death penalty equals $500e^{-2.06}/(1 + e^{-2.06})$

True

The estimated probability for a single case with a white victim and defendant is defined as $e^{-2.06}/(1+e^{-2.06})$. Then, the fitted count for 500 cases would simply be the estimated probability multiplied by 500.

Appendix

1 Donner Party

A Fit a logistic regression model for survival as a function of age. Write out the fitted model.

i Interpret the coefficient of the estimated effect of age. Do the odds of survival increase or decrease with age?

Here are the coefficient results from our GLM:

```
model1_summary = summary(model1)
model1_summary$coefficients
```

ii Determine the probability of survival for a newborn (age = 0).

```
prob_survival = predict(model1,newdata=data.frame(age=0), type="response")
```

iii Compute a 95% confidence interval for the odds ratio for age. What is the formal interpretation of this confidence interval?

```
beta_age= model1$coefficients[2]
ste_age= summary(model1)$coefficients[2,2]
OR_age = exp(beta_age)
#Confidence interval
z_star = 1.96
L = exp(beta_age - z_star*ste_age)
U = exp(beta_age + z_star*ste_age)
```

The odds for suvival multiply by $e^{-.06647}$ for every 1 unit increase in age, thus giving the result:

```
cat("\nChange in Odds From a One Unit Increase in Age = ",OR_age,"")
```

```
cat("\nConfidence Interval = [",L," , ",U,"]")
```

iv Conduct a test of lack of fit for this model based on deviance. Be sure to specify your null and alternative hypotheses, test statistics, distribution of the test statistic under the null, p-value, and your conclusions.

```
pval = pchisq(model1$null.deviance - model1$deviance, df= model1$df.null - model1$df.residual, lower.ta
```

B Consider an additive model with the variables age and sex. Fit a logistic regression model for survival as a function of age and sex. Write out the fitted model.

```
model2 = glm(survive~age+gender,family=binomial("logit"))
```

i Interpret the coefficient of the estimated effect of gender.

Here are the coefficient results for GLM considering Age and Gender as predictors:

```
model2_summary =summary(model2)
model2_summary$coefficients
```

ii Determine the probability of survival for a 24 year old male.

P(survival|age = 24, gender = Male) = $\frac{exp(3.231-0.08\times24-1.59\times1)}{1+exp(3.231-0.08\times24-1.59\times1)}$

```
prob_survival_24_m = predict(model2,newdata=data.frame(age=24,gender=1), type="response")
```

iii Determine the probability of survival for a 24 year old female.

P(survival|age = 24, gender = Female) = $\frac{exp(3.231-0.08\times24)}{1+exp(3.231-0.08\times24)}$

```
prob_survival_24_f = predict(model2,newdata=data.frame(age=24,gender=0), type="response")
```

iv Conduct a test to determine if the variable sex should be included in the model.

```
pval = pchisq(model1$deviance - model2$deviance, df=model1$df.residual - model2$df.residual,
              lower.tail=FALSE)
```

v What are the odds of survival for a 50 year old female to a 20 year old female?

The odds of survival for a 50 year old female to a 20 year old female is defined as $e^{0.07820\times(50-20)} = e^{0.07820\times30}$, giving the following

```
cat("odds of survival for a 50 year old female to a 20 year old female = ",
    exp(model2$coefficients[2]*(50-20)),"")
```

vi Write two models: one model for males and one model for females. Plot the model for males and the model for females on the same graph. Based on these models, are women more likely to survive than men?}

Here are the two probability models for each respective gender:

Model for males: $\pi = \frac{exp\left((3.23041-1.59729)-0.07820\times\mathbf{age}\right)}{1+exp\left((3.23041-1.59729)-0.07820\times\mathbf{age}\right)}$

Model for females: $\pi = \frac{exp\left(3.23041-0.07820\times\mathbf{age}\right)}{1+exp\left(3.23041-0.07820\times\mathbf{age}\right)}$

Here are the plots for our two models:

```
age.seq = seq(0,65,length=65)
Male = predict(model2,newdata=data.frame(age=age.seq,gender=1),
               type="response")
Female = predict(model2,newdata=data.frame(age=age.seq,gender=0),
                 type="response")

plot(age.seq, Male,ylim=c(0,1),type="l",col="Red",
     xlab="Age", ylab="P(Survival)",
     main="Probability of Survival by Gender")

lines(age.seq,Female,type="l",col="Blue")
```

```r
legend("topright", legend=c("Male", "Female"),
       col=c("Red", "Blue"),lty=1, cex=0.8)
```

Based on this plot, we can see that females have a consistently higher probability of survival when compared to men.

C Finally, test for the interaction between age and gender by fitting a logistic regression model for survival as a function of age and sex. Write out the fitted model. Be sure to specify your null and alternative hypotheses, test statistics, distribution of the test statistic under the null, p-value, and your conclusions. Plot the estimated model. What are your conclusions?

```r
model3=glm(survive~age+gender+age*gender,family=binomial("logit"))
```

```r
model3_summary = summary(model3)
model3_summary$coefficients
```

When considering gender, we have the following probability models:

Model for males: $\pi = \frac{exp\left(.318-.032\times\textbf{age}\right)}{1+exp\left(.318-0.032\times\textbf{age}\right)}$

Model for females: $\pi = \frac{exp\left(7.24-.194\times\textbf{age}\right)}{1+exp\left(7.24-.194\times\textbf{age}\right)}$

Using these models, here are the interaction model plots given age:

```r
age.seq = seq(0,65,length=65)
Male = predict(model3,newdata=data.frame(age=age.seq,gender=1),
               type="response")
Female = predict(model3,newdata=data.frame(age=age.seq,gender=0),
                 type="response")

plot(age.seq, Male,ylim=c(0,1),type="l",col="Red",
     xlab="Age", ylab="P(Survival)",
     main="Probability of Survival by Gender")

lines(age.seq,Female,type="l",col="Blue")

legend("topright", legend=c("Male", "Female"),
       col=c("Red", "Blue"),lty=1, cex=0.8)
```

```r
pval = pchisq(model2$deviance - model3$deviance, df=model2$df.residual - model3$df.residual,
              lower.tail=FALSE)
```

2 Book Problems

4.3

In the first nine decades of the twentieth century in baseball's National League, the percentage of times the starting pitcher pitched a complete game were: 72.7 (1900-1909), 63.4, 50.0, 44.3, 41.6, 32.8, 27.2, 22.5, 13.3 (1980-1989) (Source: George Will, Newsweek, April 10, 1989).

B

Substituting x = 12, predict the percentage of complete games for 2010 - 2019. Is this prediction plausible? Why?

```
pi = function(x){
  .7578 - .0694*x}
```

C

The logistic regression ML fit is $\hat{pi}$ = exp(1.148 - 0.315 x)/[1 + exp(1.148 - 0.315 x)]. Obtain $\hat{\pi}$ for x = 12. Is this more plausible than the prediction in (b)?

```
log.pi = function(x){
  (exp(1.148 - .315*x))/(1 + exp(1.148 - .315*x))}
```

4.4

B

Estimate the probabilities of heart disease at snoring levels 0 and 5.

```
prob.snore = function(x){
  logit = -3.866 + .397*x
  prob = exp(logit)/(1+exp(logit))
  return(prob)
}
```