Homework 3

Tyler Poelking

2/2/2017

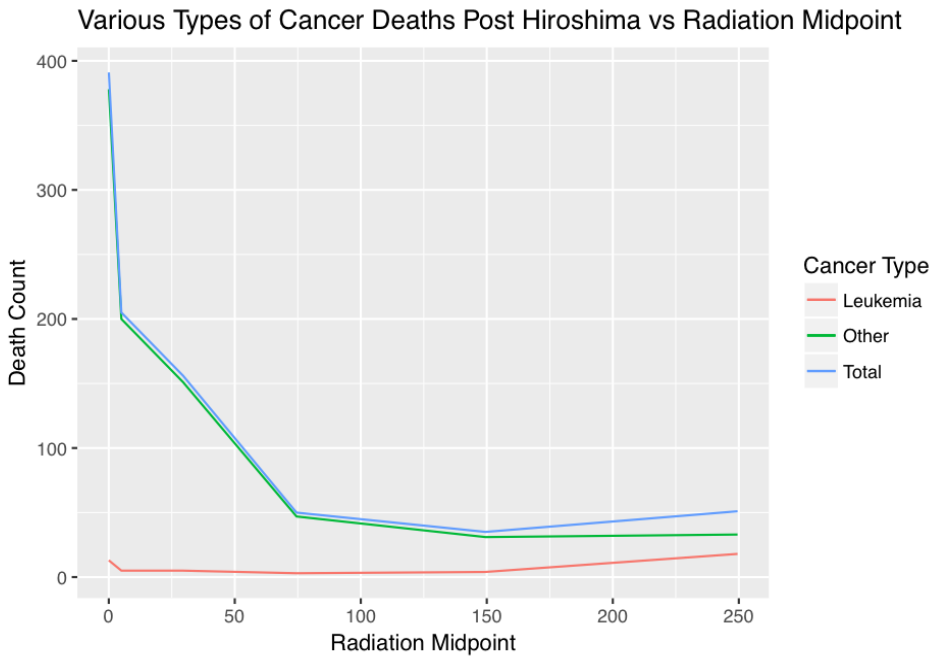
Question 1a

```
#Import
library(readr)
library(ggplot2)
hiro <- read_delim("http://www.stat.osu.edu/~pfc/teaching/3302/datasets/hiroshima.txt",
    " ", escape_double = FALSE, trim_ws = TRUE)
```

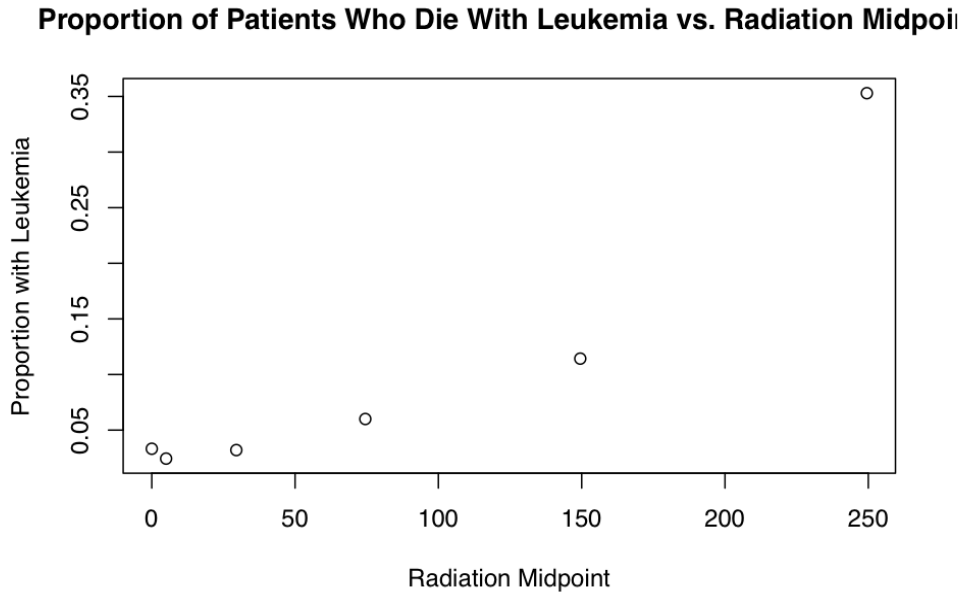
```
## Parsed with column specification:
## cols(
##   radiation = col_character(),
##   midpoint = col_double(),
##   leukemia = col_integer(),
##   other = col_integer(),
##   total = col_integer()
## )
```

```
## Warning: 1 parsing failure.
## row col  expected    actual
##   6  -- 5 columns 6 columns
```

```
plot = ggplot() +
geom_line(data = hiro, aes(x = hiro$midpoint, y = leukemia, color = "Leukemia")) +
geom_line(data = hiro, aes(x = midpoint, y = other, color = "Other"))+
geom_line(data = hiro, aes(x = midpoint, y = total, color = "Total")) +
labs(color="Cancer Type", title="Various Types of Cancer Deaths Post Hiroshima vs Radiation Midpoint") +
xlab('Radiation Midpoint') +
ylab('Death Count')
plot
```




```
#define proportion killed w leuk
p = hiro$leukemia/hiro$total
plot(x=hiro$midpoint, y=p, xlab = "Radiation Midpoint", ylab =
         "Proportion with Leukemia",
     main="Proportion of Patients Who Die With Leukemia vs. Radiation Midpoint" )
```



First plot: Patient deaths with types of cancers other than leukemia respond differently to an increase in the radiation midpoint than patients with leukemia. For other cancer types, as the radiation midpoint increases, the death count decreases dramatically at first then levels out after a midpoint radiation of approximately 100. Patient deaths with leukemia, however, respond almost in almost the exact opposite. From the plot, one can observe that as the radiation midpoint increases, the number of deaths of leukemia patients actually increases at an increasing rate.The line representing total here serves as an average of the two other lines. The abnormal response of leukemia causes the total line to deviate from the other, especially toward higher values of the radiation midpoint, where the death count for Leukemia patients rises while the line representing other cancers stays flat.

Second plot: It appears that as the radiation midpoint increases, so does the proportion of patients who died having leukemia. The only exception to this trend is when radiation midpoint went from 0 to 5. After that, the proportion of patients who died having leaukemia compared to other cancer types increases at an increasing rate. This log like patter suggestest we use a logistic model.

1.b

```
hiro = cbind(p,hiro)
hiro.logit.1 =  glm(p ~ midpoint, weight = total, family = binomial, data=hiro)

summary(hiro.logit.1)
```

```
##
## Call:
## glm(formula = p ~ midpoint, family = binomial, data = hiro, weights = total)
##
## Deviance Residuals:
##        1         2         3         4         5         6
##  0.67399  -0.41184  -0.41877  -0.08743  -0.42526   0.20237
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -3.565875   0.212254 -16.800  < 2e-16 ***
## midpoint     0.011624   0.001487   7.819 5.31e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 54.3509  on 5  degrees of freedom
## Residual deviance:  1.0287  on 4  degrees of freedom
## AIC: 26.694
##
## Number of Fisher Scoring iterations: 4
```

Looking at the coefficion summary, both the intercept and slope parameters are significantly different from zero at the a=0.05 level. This may suggest that the radiation midpoint is related to the probability the patient dead has leukemia. Based on the summary of this model, we can see that, as the radiation midpoint increase by one unit the expected value of the logit of the proportion of patients dead with leukemia increased by 0.011624. For example, increasing the radiation midpoint from 29.5 to 74.5 increases the logit of the proportion dead by 45* 0.011624 = 0.52308. We can also say that the multiplicative change in odds for a unit increase in radiation midpoint is e^0.011624 = 1.01169. The intercept of the logit of the proporiton of dead patients who had leukemia is -3.565875. This can be interpreted as the logit of the expected portion of dead patients who had leukemia with a midpoint radiation value of 0.

```
anova(hiro.logit.l, test="Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: p
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         5     54.351
## midpoint  1   53.322         4      1.029 2.831e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova table above compares the intercept only model(NULL) with the simple logistic regression model (midpoint). The degree of freedom change when going from the NULL to SLR model is 1, since the radiation midpoint is added. The decrease in residual deviance from the NULL to the model with the midpoint is 53.322 and serves as the test statistic that determines whether or not adding the radiation midpoint will increse the predictive value of our model. The Ho under this test is the NULL model while the Ha is the SLR model with the added parameter. Since the p-value for this hypothesis test is so small, we determine that adding midpoint makes our model significantly stronger.

1.c

```
mp = (50+99)/2


pred.logit= predict(hiro.logit.l, data.frame(midpoint = mp), se.fit = T)

the.ci.for.logit = c(pred.logit$fit - 1.96*pred.logit$se.fit, pred.logit$fit +1.96*pred.logit$se.fit)
round(the.ci.for.logit, 3)
```

```
##      1      1
## -3.014 -2.386
```

```
inv.logit(the.ci.for.logit)
```

```
##          1          1
## 0.04679794 0.08426644
```

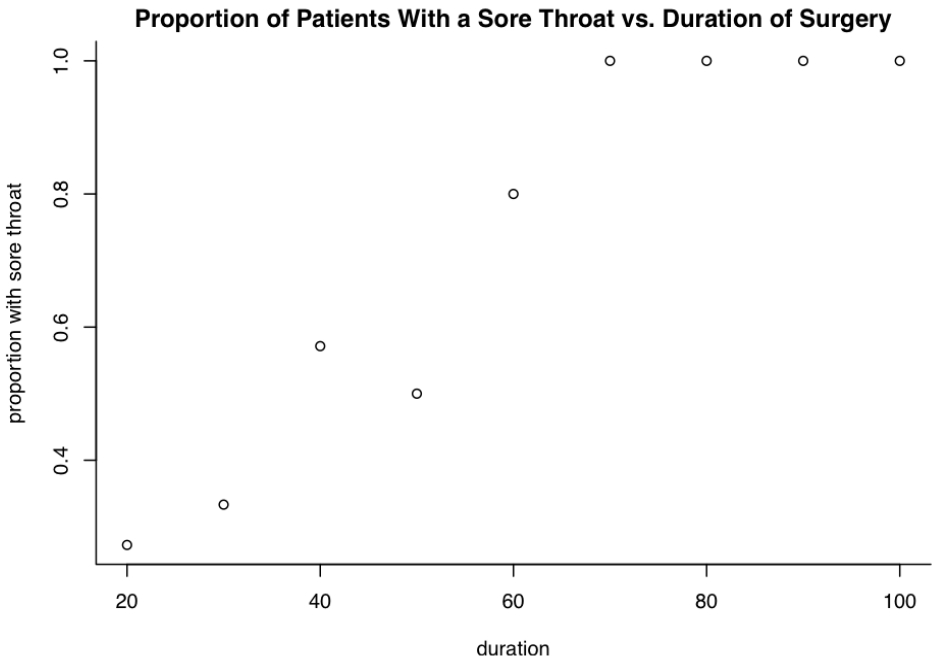
A 95% confidence interval for the logit probability of having leukemia when the dose is between 50 and 99 radons is (-3.014, -2.386). Usingthe inv.logit function, this translates to a 95% confidence interval that the probability of having leaukemia when the dose is between 50 and 99 radons is (0.04679794, 0.08426644). Upon repeating this experiment, we can expect the true proportion of patients having leaukemia for this radiation dosage to fall within the produced confidence interval 95% of the time.

Question 2.a and 2.b

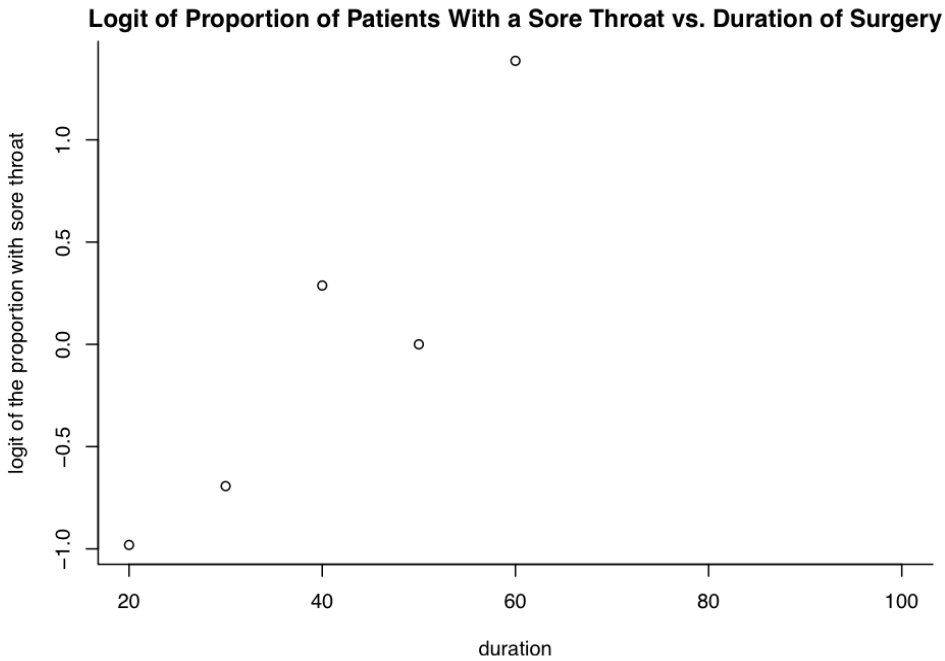
```
#read in data
p_throats <- read_delim("http://www.stat.osu.edu/~pfc/teaching/3302/datasets/throat.txt",
                        "\t", escape_double = FALSE, trim_ws = TRUE)
```

```
## Parsed with column specification:
## cols(
##   throat = col_integer(),
##   duration = col_integer()
## )
```

```
#Round duration
p_throats$duration= round(p_throats$duration, -1)

## Here are the total number of patients with sore throats by the duration
total <- table(p_throats$duration)

## Here are the number of patients with sore throats(1) or without (0), by duration
sore.or.not <- table(p_throats$duration, p_throats$throat)

# Summarize the number sore or not sore by the duration.
sore.or.not = cbind(sore.or.not, total = total)


## calculate the proportion killed.
prop.sore <- as.numeric(sore.or.not[,2] / total)

unique.duration <- sort(unique(p_throats$duration))

## Produce a plot of the duration versus the proportion and the logit
par(mfrow=c(1,1), cex=0.8, mar=c(4.1, 4.1, 2, 1), bty="L")

plot(unique.duration, prop.sore,
     xlab="duration", ylab="proportion with sore throat",
     main="Proportion of Patients With a Sore Throat vs. Duration of Surgery")
```




```
plot(unique.duration, logit(prop.sore),
     xlab="duration", ylab="logit of the proportion with sore throat",
     main="Logit of Proportion of Patients With a Sore Throat vs. Duration of Surgery")
```



Based on the two charts above, the duration of surgery appears to have a significant impact on the proportion of patients with a sore throat. Particularly, as the duration of the surgery increases, the proportion of patients with a sore throat tends to rise. In the first plot, the postive correlation of this relationship increases as duration increases. This is especially noticeable by the fact that the proportion of patients with a sore throat is 1.0 for surgery durations 70, 80, 90, and 100. In the second plot, where the logit is on the y axis, this relationship appears much more linear, at least for the first 5 duration values. Since the proportion is 1.0 for the surgurery durations mentioned above, the logit function outputs a "inf", due to a zero in the denominator. Thus, these points are not plotted.

2.c The simple logistic regression model the predicts the probability of a patient that has surgery having a sore through using surgery duration is logit(p) = B0 + B1*duration, where p is the probability a patient that has surgery has a sore throat and duration is, of course, the surgery duration.

2.d

```
p_throats.logit.l =  glm(throat~duration, family = binomial, data=p_throats)

summary(p_throats.logit.l)
```

```
##
## Call:
## glm(formula = throat ~ duration, family = binomial, data = p_throats)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8725  -0.7496   0.2803   0.7761   1.6773
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -2.47004    1.01303  -2.438  0.01476 *
## duration     0.06722    0.02513   2.675  0.00748 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 46.662  on 33  degrees of freedom
## Residual deviance: 35.154  on 32  degrees of freedom
## AIC: 39.154
##
## Number of Fisher Scoring iterations: 5
```

Looking at the coeffiction summary, both the intercept and slope parameters are significantly different from zero at the a=0.05 level. This may suggest that the surgery duration is related to the probability the patient has a sore throat. Based on the summary of this model, we can see that, as the surgery duration increase by one unit the expected value of the logit of the proportion of patients with sore throats increases by 0.011624. For example, increasing the surgery duration from 20 to 30 increases the logit of the proportion with sore throats by 10* 0.06722 = 0.6722. We can also say that the multiplicative change in odds for a unit increase in surgery duration is eˆ0.06722 = 1.06953. The intercept of the logit of the proportion of patients with sore throats is -2.47004. This does not have any contextual meaning to the problem because we cannot have a surgery whose duration was 0.

```
anova(p_throats.logit.l, test="Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: throat
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                        33     46.662
## duration  1   11.509        32     35.154 0.0006927 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova table above compares the intercept only model(NULL) with the simple logistic regression model (duration). The degree of freedom change when going from the NULL to SLR model is 1, since the radiation duration is added. The decrease in residual deviance from the NULL to the model with the duration is 11.509 and serves as the test statistic that determines whether or not adding the surgery duration will increse the predictive value of our model. The Ho under this test is the NULL model while the Ha is the SLR model with the added parameter. Since the p-value for this hypothesis test is so small, we determine that adding duration makes our model significantly stronger.

2.e

```
d = 30

zstat = qnorm(1-(0.01/2))

pred.logit= predict(p_throats.logit.l, data.frame(duration = d), se.fit = T)

the.ci.for.logit = c(pred.logit$fit - zstat*pred.logit$se.fit, pred.logit$fit +zstat*pred.logit$se.fit)
```

```
round(the.ci.for.logit, 3)
```

```
##      1      1
## -1.604  0.697
```

```
inv.logit(the.ci.for.logit)
```

```
##         1         1
## 0.1674455 0.6674731
```

A 99% confidence interval for the logit probability of having a sore throat when the duration is 30 (-1.604, 0.697). Usingthe inv.logit function, this translates to a 99% confidence interval that the probability of having leaukemia when the dose is between 50 and 99 radons is (0.1674455 0.6674731). Upon repeating this experiment, we can expect the true proportion of patients having leaukemia for this radiation dosage to fall within the produced confidence interval 99% of the time.

2.f

Yes, underlying factors that determine whether a patient received a sore throat could be devoted to aspects of the hospital they are at. Changing hospitals implies changing doctors, nurses, materials, surgical methods used, cleanliness of the environments, etc, and the proportion of patients who get sore throats for any particular surgical duration could very easily differ. Basically, the experiment would be less controlled and data not as independent, thus compromising our confidence to make inferences on the proportion of patients with sore throats based solely on surgery duration.