HW 4 Solutions

1. Using age group as predictor variable

a. Two-way table of age group by survival status:

```
xtabs(~AgeGroup+Survive,data=ICU)
```

```
##          Survive
## AgeGroup  0  1
##        1  5 54
##        2 17 60
##        3 18 46
```

```
prop.table(xtabs(~AgeGroup+Survive,data=ICU), 1) #proportion table of survival within each age group
```

```
##          Survive
## AgeGroup          0          1
##        1 0.08474576 0.91525424
##        2 0.22077922 0.77922078
##        3 0.28125000 0.71875000
```

b. There's a clear relationship between age group and survival, with young adults (group 1) surviving 91.5% of the time, middle-age adults surviving 78% of the time, and older adults (group 3) surviving 72% of the time.

c. Odds of survival for each group:

```
xtabs(~AgeGroup+Survive,data=ICU)
```

```
##          Survive
## AgeGroup  0  1
##        1  5 54
##        2 17 60
##        3 18 46
```

```
xtabs(~AgeGroup+Survive,data=ICU)[,2]/ xtabs(~AgeGroup+Survive,data=ICU)[,1] #odds of survival within e
```

```
##         1         2         3
## 10.800000  3.529412  2.555556
```

d. OR for young people compared to middle-aged people = (odds for group 1)/(odds for group 2) = 10.8/3.52 = 3.06. Young people have slightly over 3 times the odds of surviving the ICU, compared to middle-aged people.

e.

```
age.group <- glm(Survive ~ as.factor(AgeGroup), data=ICU, family=binomial); summary(age.group)
```

```
##
## Call:
## glm(formula = Survive ~ as.factor(AgeGroup), family = binomial,
##     data = ICU)
##
```

```
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.2218   0.4208   0.7063   0.7063   0.8127  
## 
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)    
## (Intercept)            2.3795     0.4675   5.090 3.57e-07 ***
## as.factor(AgeGroup)2  -1.1184     0.5422  -2.063  0.03915 *  
## as.factor(AgeGroup)3  -1.4413     0.5439  -2.650  0.00805 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 191.59  on 197  degrees of freedom
## AIC: 197.59
## 
## Number of Fisher Scoring iterations: 5
```

f. The intercept is the "fitted model" when age group = 1. So when age group = 1, we can write the model as:

log(odds of survival) = 2.38

Thus, we can say that the odds of survival for young people in the ICU is eˆ(2.38) = 10.8 (which we already found in part c).

g. The eˆ(slope) for any group is the **odds ratio** of that group compared the the "reference" group, which is age group 1 in this case. So, eˆ(slope) for age group 2 is the odds ratio of middle-age people (group 2) to young people (group 1). This means that middle-age people have 33% the odds of survival of young people in the ICU.

h. 0.33 = 1/3.06, which makes sense since we flipped the order of comparison between part (d) and part (g).

i. The drop in deviance test is below. The test stat is 8.57 and the p-value is 0.014, which is strong evidence that age group is a statistically significant predictor of survival in the ICU.

```
anova(age.group, test="Chisq")
```

```
## Analysis of Deviance Table
## 
## Model: binomial, link: logit
## 
## Response: Survive
## 
## Terms added sequentially (first to last)
## 
## 
##                     Df Deviance Resid. Df Resid. Dev Pr(>Chi)  
## NULL                                  199     200.16           
## as.factor(AgeGroup)  2   8.5721       197     191.59  0.01376 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

j. From the second line of code below, we see that a 90% CI on eˆ(slope) for Age group 3 is 0.09 to 0.55. This means that we are 90% confident that the odds of an older person surviving the ICU are between 9% and 55% the odds of a young person surviving.

```
confint(age.group, level=0.9) #CI for slopes
```

```
## Waiting for profiling to be done...

##                              5 %       95 %
## (Intercept)            1.682012  3.2430992
## as.factor(AgeGroup)2 -2.079850 -0.2715962
## as.factor(AgeGroup)3 -2.405642 -0.5926046
```

```
exp(confint(age.group, level=0.9)) #CI for e^(slope), which is what we want
```

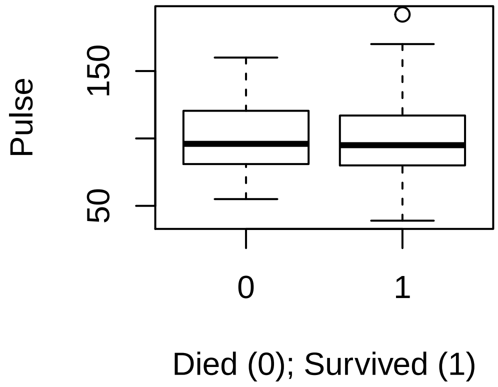
```
## Waiting for profiling to be done...

##                               5 %       95 %
## (Intercept)           5.37635985 25.6129774
## as.factor(AgeGroup)2 0.12494898  0.7621619
## as.factor(AgeGroup)3 0.09020755  0.5528854
```

k. Linearity is automatic because age group is a categorical variable. A plot of the empirical logits vs. age group would draw a "line" between young people and middle-aged people, and another line between young people and old people. Since two points always make a line, linearity is met.

2. Using pulse as predictor variable

a. Here is a boxplot comparing the pulse distribution for survivors and non-survivors.

```
boxplot(Pulse~Survive,data=ICU,ylab="Pulse",xlab="Died (0); Survived (1)")
```



b. The two pulse distributions look similar, other than survivors having a wider spread. There does not appear to be a relationship between pulse and survival.

c.

```
pulse.log <- glm(Survive ~ Pulse, data=ICU, family=binomial); summary(pulse.log)
```

```
##
## Call:
## glm(formula = Survive ~ Pulse, family = binomial, data = ICU)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8524   0.6339   0.6579   0.6784   0.7533
##
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)  
## (Intercept)  1.679129   0.679863   2.470   0.0135 *
## Pulse       -0.002941   0.006552  -0.449   0.6535  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 199.96  on 198  degrees of freedom
## AIC: 203.96
##
## Number of Fisher Scoring iterations: 4
```

d. e^(slope) = 0.997. This means that with a 1-unit increase in pulse, we expect odds of survival to decrease by 0.3%.

```r
exp(-0.002941)
```

```
## [1] 0.9970633
```

e. No, pulse is not a statistically significant predictor of survival, as shown by the Wald test (test stat = -0.449, p-value = 0.6535).

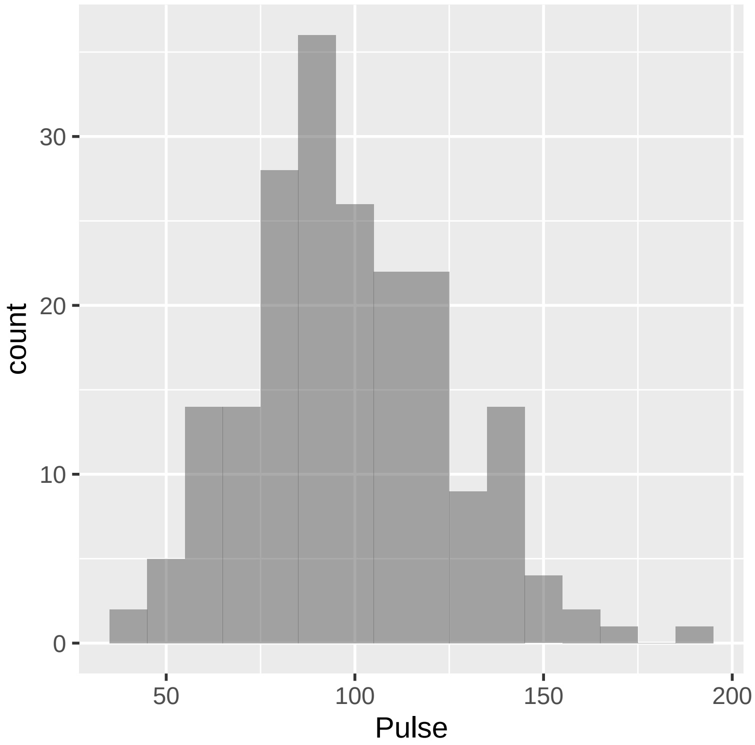
f. To check linearity I need to make an empirical logit plot, which means I need to do the grouping thing that we did in “Odds & ORs, Part 2”.

First, I need to create groups of about the same size.

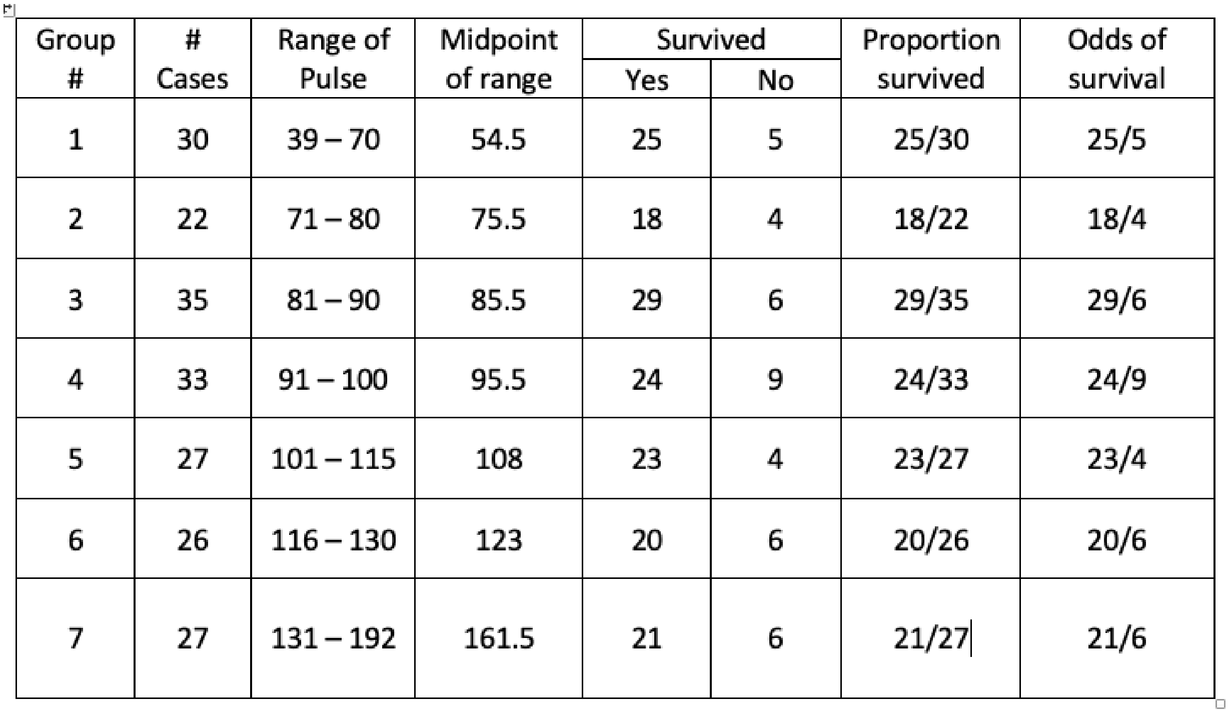
```r
xtabs(~Pulse,data=ICU)
```

```
## Pulse
##  39  44  46  48  52  55  58  59  60  62  64  65  66  67  68  70  71  72  73  74 
##   1   1   1   1   2   1   1   1   6   1   3   2   2   1   2   4   1   1   1   1 
##  75  76  78  79  80  81  83  84  85  86  87  88  89  90  91  92  94  95  96  98 
##   1   2   4   2   9   3   3   3   2   6   1   9   2   6   1   6   2   3   6   2 
##  99 100 103 104 105 106 107 108 109 110 111 112 114 115 116 118 119 120 121 122 
##   4   9   3   1   1   3   1   2   1   3   1   6   3   2   1   2   1   7   1   2 
## 124 125 126 128 131 132 135 136 137 138 140 143 144 145 150 153 154 160 162 170 
##   4   4   1   3   1   2   2   2   1   1   6   1   1   2   2   1   1   1   1   1 
## 192 
##   1
```

```r
gf_histogram(~Pulse, data=ICU, binwidth = 10)
```



Here is the summary table I made:



Group #	# Cases	Range of Pulse	Midpoint of range	Survived		Proportion survived	Odds of survival
				Yes	No		
1	30	39 – 70	54.5	25	5	25/30	25/5
2	22	71 – 80	75.5	18	4	18/22	18/4
3	35	81 – 90	85.5	29	6	29/35	29/6
4	33	91 – 100	95.5	24	9	24/33	24/9
5	27	101 – 115	108	23	4	23/27	23/4
6	26	116 – 130	123	20	6	20/26	20/6
7	27	131 – 192	161.5	21	6	21/27	21/6

(Here's examples of the code I used to find these counts. There are lots of other ways to do this.)

```
count(ICU$Pulse>70 & ICU$Pulse<91) #counting how many cases are in a certain range
```

```
## n_TRUE
##     57
```

```
tally(~Survive, data=ICU[ICU$Pulse>70 & ICU$Pulse<81,]) #tallying the survivals & deaths in that range
```

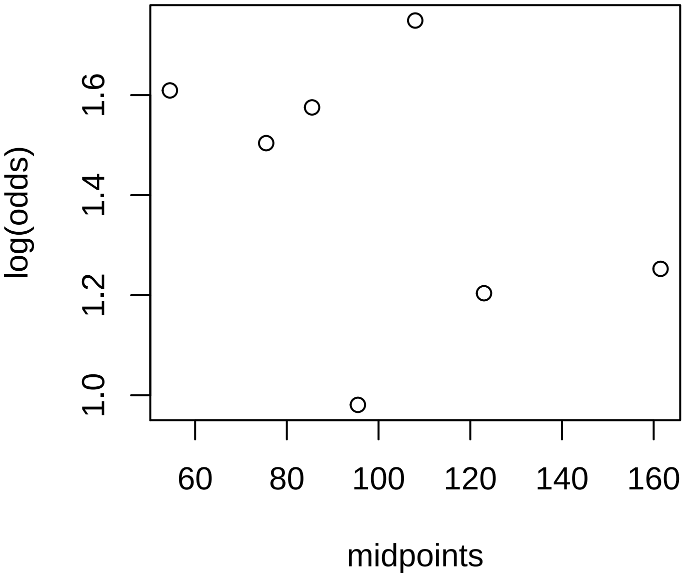
```
## Survive
##  0  1
##  4 18
```

Now we create midpoint and odds vectors to plot:

```
midpoints <- c(54.5, 75.5, 85.5, 95.5, 108, 123, 161.5)
odds <- c(25/5, 18/4, 29/6, 24/9, 23/4, 20/6, 21/6)
```

Plot log(odds) vs midpoints. We want this to be linear!

```
plot(log(odds)~midpoints)
```



I don't really see linearity here. As I've said before, the group process is not perfect: are introducing bias by choosing these groups. So it's hard to say if this is really non-linear or just poorly-chosen groups? The good news is: it kind of doesn't matter, since Pulse is not a significant predictor anyway!

3. Using infection as predictor variable

a.

```
infection.log <- glm(Survive ~ Infection, data=ICU, family=binomial); summary(infection.log)
```

```
##
## Call:
## glm(formula = Survive ~ Infection, family = binomial, data = ICU)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9905   0.5448   0.5448   0.8203   0.8203
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.8326     0.2693   6.806    1e-11 ***
## Infection    -0.9163     0.3617  -2.533   0.0113 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 193.59  on 198  degrees of freedom
## AIC: 197.59
##
## Number of Fisher Scoring iterations: 4
```

b. e^(slope) = 0.4, which means that those who have an infection have 40% the odds of survival of those who don't have an infection.

```
exp(-0.9163)
```

```
## [1] 0.3999963
```

c. According to the Wald test (test stat=-2.533, p-value=0.0113), there is strong evidence for Infection as a statistically significant predictor of survival in the ICU.

d. We are 95% confident that the odds of someone with an infection surviving the ICU are between 19% and 81% the odds of someone with no infection surviving.

```
exp(confint(infection.log))
```

```
## Waiting for profiling to be done...
```

```
##                2.5 %     97.5 %
## (Intercept) 3.800364 10.9989452
## Infection   0.193822  0.8064948
```

e. Linearity is automatically met because this is a binary predictor.

4. Independence & Randomness

Independence: In order to answer the question about independence, I would like more information about these patients. They were all in one hospital's ICU, but were they there at the same time? Were they being treated by the same staff? I can imagine that two patients' outcomes might be dependent if they were being treated by an (for example) an incompetent staff member, or if an infection swept through the hospital. But overall, there results are probably reasonably independent.

Randomness: It does not state anything about these patients being randomly sampled, but it does say that they are "a sample of 200 patients who were part of a larger study conducted in a hospital's Intensive Care Unit". Can we treat them as randomly selected? Is there a reason to believe that they differ systematically from other patients at this hospital's ICU? We don't really have enough information to answer these questions.