HW 5 Solutions

1. Using age group and presence of infection

a.

```
model1 <- glm(Survive ~ as.factor(AgeGroup)+Infection, data=ICU, family=binomial); summary(model1)
```

```
##
## Call:
## glm(formula = Survive ~ as.factor(AgeGroup) + Infection, family = binomial,
##     data = ICU)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3504   0.3612   0.5743   0.6826   0.9706
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            2.6968     0.5008   5.385 7.26e-08 ***
## as.factor(AgeGroup)2  -0.9780     0.5497  -1.779   0.0752 .
## as.factor(AgeGroup)3  -1.3588     0.5497  -2.472   0.0134 *
## Infection             -0.8300     0.3701  -2.243   0.0249 *
## ---
## 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: 186.45  on 196  degrees of freedom
## AIC: 194.45
##
## Number of Fisher Scoring iterations: 5
```

b. The p-value of 0.0134 is telling us that there is a signficant difference in odds of survival between older people (AgeGroup3) and younger people (AgeGroup1), even when controlling for infection.

c. The intercept is the "fitted model" when AgeGroup=1 and Infection=0. So when when AgeGroup=1 and Infection=0, we can write the model as:

log(odds of survival) = 2.697

Thus, we can say that the odds of survival for young people in the ICU who do NOT have signs of infection is e^(2.697) = 14.8. Thus, for each young person without infection who dies in the ICU, we expect nearly 15 to survive.

d. The e^(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, e^(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 37.6% the odds of survival of young people in the ICU, assuming the same infection status.

e. From the second line of code below, we see that a 90% CI on e^(slope) for Age group 3 is 0.097 to 0.607.

This means that we are 90% confident that the odds of an older person surviving the ICU are between 10% and 61% the odds of a young person surviving, assuming the same infection status.

```
confint(model1, level=0.9) #CI for slopes
```

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

```
##                              5 %       95 %
## (Intercept)            1.942832  3.6103641
## as.factor(AgeGroup)2 -1.949229 -0.1162658
## as.factor(AgeGroup)3 -2.330928 -0.4989041
## Infection            -1.448758 -0.2266704
```

```
exp(confint(model1, level=0.9)) #CI for e^(slope), which is what we want
```

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

```
##                               5 %       95 %
## (Intercept)           6.97848313 36.9795158
## as.factor(AgeGroup)2 0.14238380  0.8902386
## as.factor(AgeGroup)3 0.09720551  0.6071957
## Infection            0.23486186  0.7971835
```

f. Linearity is automatic because age group and infection are both categorical variables. 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. Adding sex

a. Two-way table of survival status by sex:

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

```
##    Survive
## Sex   0   1
##   0  24 100
##   1  16  60
```

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

```
##    Survive
## Sex         0         1
##   0 0.1935484 0.8064516
##   1 0.2105263 0.7894737
```

b. There does not appear to be a relationship between survival and sex. 80.6% of males survive and 78.9% of females.

c.

```
model2 <- glm(Survive ~ as.factor(AgeGroup)+Infection+Sex, data=ICU, family=binomial); summary(model2)
```

```
##
## Call:
## glm(formula = Survive ~ as.factor(AgeGroup) + Infection + Sex,
##     family = binomial, data = ICU)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.3521   0.3604   0.5731   0.6847   0.9732
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.70128    0.51795   5.215 1.83e-07 ***
## as.factor(AgeGroup)2 -0.97805    0.54973  -1.779   0.0752 .
## as.factor(AgeGroup)3 -1.35734    0.55125  -2.462   0.0138 *
## Infection            -0.82996    0.37009  -2.243   0.0249 *
## Sex                  -0.01273    0.37659  -0.034   0.9730
## ---
## 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: 186.45  on 195  degrees of freedom
## AIC: 196.45
##
## Number of Fisher Scoring iterations: 5
```

d. According to the Wald test (test stat=-0.034, p-value=0.9730), there is no evidence for Sex as a statistically significant predictor of survival in the ICU, once AgeGroup and Infection are already in the model.

3. Adding emergency

a. Two-way table of survival status by emergency:

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

```
##           Survive
## Emergency   0   1
##         0   2  51
##         1  38 109
```

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

```
##           Survive
## Emergency          0          1
##         0 0.03773585 0.96226415
##         1 0.25850340 0.74149660
```

b. There does appear to be a relationship between survival and sex. 96.2% of elective admissions survive, but only and 74.1% of emergency admissions.

c.

```
model3 <- glm(Survive ~ as.factor(AgeGroup)+Infection+Emergency, data=ICU, family=binomial); summary(mo
```

```
##
## Call:
## glm(formula = Survive ~ as.factor(AgeGroup) + Infection + Emergency,
##     family = binomial, data = ICU)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4958   0.2455   0.4131   0.7561   1.0941
##
```

```
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            4.7995     0.8796   5.456 4.86e-08 ***
## as.factor(AgeGroup)2  -1.3123     0.5613  -2.338  0.01939 *
## as.factor(AgeGroup)3  -1.7304     0.5690  -3.041  0.00236 **
## Infection             -0.4887     0.3923  -1.246  0.21283
## Emergency             -2.3810     0.7604  -3.131  0.00174 **
## ---
## 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: 169.60  on 195  degrees of freedom
## AIC: 179.6
##
## Number of Fisher Scoring iterations: 6
```

d. According to the Wald test (test stat=-3.13, p-value=0.0017), there is strong evidence for Emergency as a statistically significant predictor of survival in the ICU, once AgeGroup and Infection are already in the model.

e. e^(slope) = 0.09, which means that those who were admitted through the ER have 9% the odds of survival of those with elective admission.

```
exp(-2.3810)
```

```
## [1] 0.09245807
```

4. Deleting Infection

a. The p-value of the nested drop-in-deviance test is 0.212, which indicates that the reduced model (model4) is sufficient. That is, Infection is not a useful addition to the model containinng AgeGroup and Emergency.

```
model4 <- glm(Survive ~ as.factor(AgeGroup)+Emergency, data=ICU, family=binomial) #first, we must fit t
anova(model4, model3, test="Chisq") #nested drop-in-deviance to compare model4 to model3
```

```
## Analysis of Deviance Table
##
## Model 1: Survive ~ as.factor(AgeGroup) + Emergency
## Model 2: Survive ~ as.factor(AgeGroup) + Infection + Emergency
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       196     171.16
## 2       195     169.60  1   1.5587   0.2119
```

b.

```
summary(model4)
```

```
##
## Call:
## glm(formula = Survive ~ as.factor(AgeGroup) + Emergency, family = binomial,
##     data = ICU)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4388   0.2632   0.4469   0.8536   1.0137
```

```
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)             4.7771     0.8801   5.428  5.7e-08 ***
## as.factor(AgeGroup)2   -1.4317     0.5527  -2.590 0.009585 **
## as.factor(AgeGroup)3   -1.8557     0.5606  -3.310 0.000931 ***
## Emergency              -2.5234     0.7538  -3.347 0.000816 ***
## ---
## 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: 171.16  on 196  degrees of freedom
## AIC: 179.16
##
## Number of Fisher Scoring iterations: 6
```

c. log(odds of survival) = 4.777 - 1.432(AgeGroup=2) - 1.856(AgeGroup=3) - 2.523(Emergency=1)

d. Fitted model for middle-age person with ER admission: log(odds of survival) = 4.777 - 1.432(AgeGroup=2) - 2.523(Emergency=1)

Fitted model for middle-age person with elective admission: log(odds of survival) = 4.777 - 1.432(AgeGroup=2)

```
exp(4.777 - 1.432 - 2.523)/(1+exp(4.777 - 1.432 - 2.523)) #prob for middle-age person with ER admission
```

```
## [1] 0.6946607
```

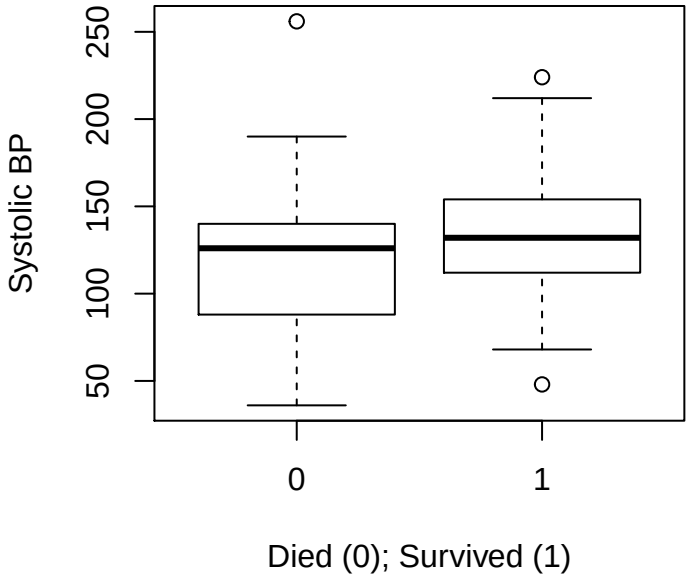
```
exp(4.777 - 1.432)/(1+exp(4.777 - 1.432)) #prob for middle-age person with ER admission
```

```
## [1] 0.9659407
```

5. Adding SysBP

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

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



b. Those who survived seem to have higher BP, on average. There might be a relationship between BP and survival.

c.

```
model5 <- glm(Survive ~ as.factor(AgeGroup)+Emergency+SysBP, data=ICU, family=binomial); summary(model5)
```

```
##
## Call:
## glm(formula = Survive ~ as.factor(AgeGroup) + Emergency + SysBP,
##     family = binomial, data = ICU)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5777   0.2024   0.4089   0.7191   1.2671
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            2.97648    1.19821   2.484  0.01299 *
## as.factor(AgeGroup)2 -1.39703    0.55821  -2.503  0.01232 *
## as.factor(AgeGroup)3 -1.83227    0.56438  -3.247  0.00117 **
## Emergency            -2.34673    0.75911  -3.091  0.00199 **
## SysBP                 0.01275    0.00595   2.142  0.03219 *
## ---
## 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: 166.19  on 195  degrees of freedom
## AIC: 176.19
##
## Number of Fisher Scoring iterations: 6
```

d. The p-value for SysBP is 0.032, which means that there is moderately strong evidence of it being a significant predictor of survival, assuming AgeGroup and Emergency are already in the model.

e. We are 95% confident that a one-unit (mm/Hg) increase in systolic BP is associated with an increase in odds of survival between 0.01% and 2.5%, assuming age group and admission location (ER or elective) stay the same.

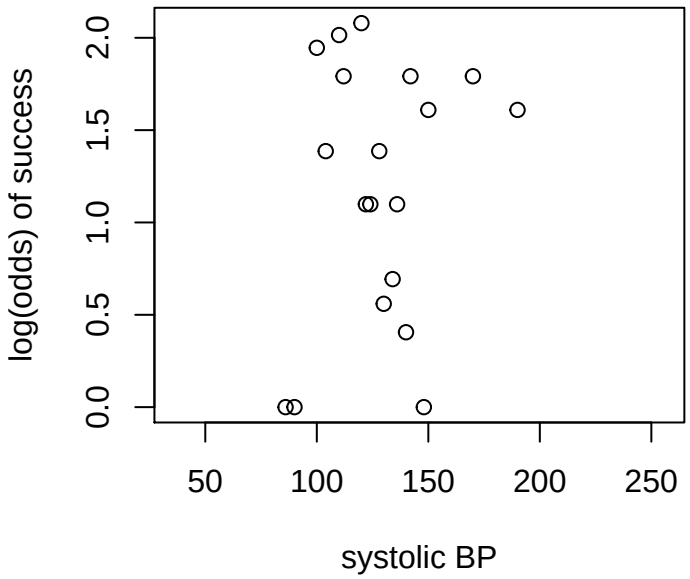
```
exp(confint(model5))
```

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

```
##                           2.5 %      97.5 %
## (Intercept)          2.04107509 244.1467138
## as.factor(AgeGroup)2 0.07511961   0.6972408
## as.factor(AgeGroup)3 0.04802032   0.4554449
## Emergency            0.01493633   0.3420080
## SysBP                1.00149488   1.0252901
```

f. To check linearity I need to make an empirical logit plot. Let's see if the "quick-and-dirty" method from Multiple Logistic Regression works here:

```
tab.sysbp <- xtabs(~SysBP+Survive,data=ICU)
prop.sysbp <- tab.sysbp[,2]/(tab.sysbp[,2] + tab.sysbp[,1])
plot(log(prop.sysbp/(1-prop.sysbp))~sort(unique(ICU$SysBP)),xlab="systolic BP",ylab="log(odds) of succes
```



Okay, so this didn't really work. The problem is that there are too many BP values with only one or two individuals, as seen in the table below. So the code above isn't helpful, because there are too many BP values with 0 "successes" or 0 "failures", which don't get included on this graph.

So check linearity I need to do the grouping thing that we did in "Odds & ORs, Part 2" and Homework 4.

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

```
xtabs(~SysBP,data=ICU)
```

```
## SysBP
##  36  48  56  62  64  66  68  70  78  80  86  90  91  92 100 104 108 110 112 114
##   1   1   1   1   1   1   1   1   1   3   2   4   1   2   8   5   2  17   7   1
## 116 118 120 122 124 126 128 130 131 132 134 135 136 138 139 140 141 142 144 146
##   2   1   9   4   4   4   5  11   1   6   3   1   4   3   1  10   1   7   3   2
## 148 150 152 154 156 158 160 162 164 168 169 170 174 180 188 190 200 204 206 208
##   4   6   2   2   3   3   4   5   1   1   1   7   2   1   1   6   2   1   1   1
## 212 224 256
##   1   1   1
```

Here is the summary table I made:

Group #	# Cases	Range of SysBP	Midpoint of range	Survived		Odds of survival
				Yes	No	
1	36	36 – 109	72.5	22	14	22/14
2	28	110 – 119	114.5	25	3	25/3
3	26	120 – 129	124.5	22	4	22/4
4	30	130 – 139	134.5	24	6	24/6
5	27	140 – 149	144.5	19	8	19/8
6	28	150 – 169	159.5	26	2	26/2
7	25	170 – 256	213	22	3	22/3

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

```
count(ICU$SysBP>109 & ICU$SysBP<120) #counting how many cases are in a certain range
```

```
## n_TRUE
##     28
```

```
tally(~Survive, data=ICU[ICU$SysBP>109 & ICU$SysBP<120,]) #tallying the survivals & deaths in that rang
```

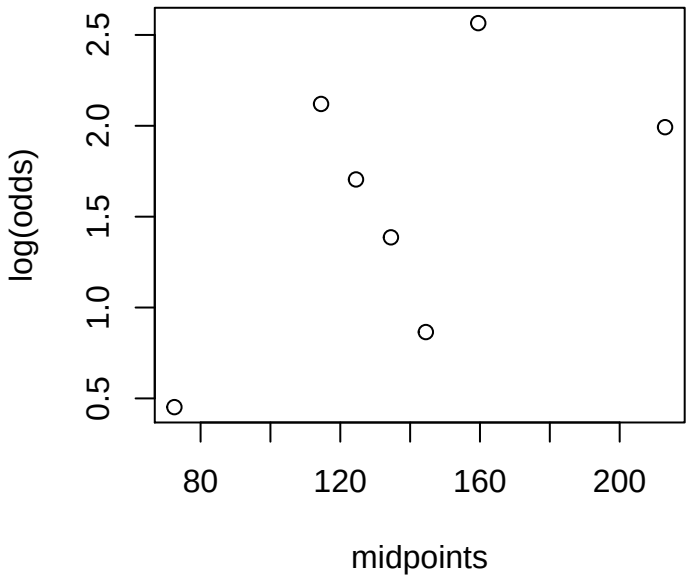
```
## Survive
##  0  1
##  3 25
```

Now we create midpoint and odds vectors to plot:

```
midpoints <- c(72.5, 114.5, 124.5, 134.5, 144.5, 159.5, 213)
odds <- c(22/14, 25/3, 22/4, 24/6, 19/8, 26/2, 22/3)
```

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

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



Wow, hard to say. As I've said before, the grouping process is not perfect: we are introducing bias by choosing these groups. So it's hard to say if this is really non-linear or just poorly-chosen groups? As we look closely, it almost seems like there's three *big* groups: really low survival for those with low BP, better survival for those with "normal" BP (between 110 and 150), and then even higher survival for those with elevated BP (150+). In that case, linearity is probably okay.

6. Comparison

a. Misclassification

```
#model 1:
pred.success.1 <- ifelse(fitted(model1)>0.5,1,0)
tally(~pred.success.1+model1$y,data=ICU, format="proportion")
```

```
##                model1$y
## pred.success.1   0   1
##              1 0.2 0.8
```

```
#model 4:
pred.success.4 <- ifelse(fitted(model4)>0.5,1,0)
tally(~pred.success.4+model4$y,data=ICU, format="proportion")
```

```
##                model4$y
## pred.success.4   0   1
##              1 0.2 0.8
```

Models 1 and 4 both have 20% misclassification. However, they predict everyone will survive!

```
#model 5:
pred.success.5 <- ifelse(fitted(model5)>0.5,1,0)
tally(~pred.success.5+model5$y,data=ICU, format="proportion")
```

```
##                model5$y
## pred.success.5     0     1
##              0 0.035 0.025
##              1 0.165 0.775
```

Model 5 has 19% misclassification; importantly, it is at least predicting that some of the individuals will die.

It may make more sense to use something other than 50% as the "threshold of success" here. Recall that 80% of the patients survived the ICU. So a threshold of 80% probably makes more sense.

```
#model 1:
pred.success.1 <- ifelse(fitted(model1)>0.8,1,0)
tally(~pred.success.1+model1$y,data=ICU, format="proportion")
```

```
##                model1$y
## pred.success.1     0     1
##              0 0.145 0.365
##              1 0.055 0.435
```

```
#model 4:
pred.success.4 <- ifelse(fitted(model4)>0.8,1,0)
tally(~pred.success.4+model4$y,data=ICU, format="proportion")
```

```
##                model4$y
## pred.success.4     0     1
##              0 0.165 0.310
##              1 0.035 0.490
```

```
#model 5:
pred.success.5 <- ifelse(fitted(model5)>0.8,1,0)
tally(~pred.success.5+model5$y,data=ICU, format="proportion")
```

```
##                model5$y
## pred.success.5     0     1
##              0 0.155 0.270
##              1 0.045 0.530
```

Now we see substantial differences, with Model 5 having much lower misclassification rates than the other two models. Model5 has the lowest rate of patients who were predicted to die but actually survived, and Model4 has the lowest % of patients who were predicted to survive but actually died.

b. Let's summarize the three models:

Model1: 2 variables, residual deviance = 186.45, misclass rate = 42%, linearity condition definitely met

Model4: 2 variables, residual deviance = 171.16, misclass rate = 34.5%, linearity condition definitely met

Model5: 3 variables, residual deviance = 166.19, misclass rate = 31.5%, linearity condition unclear

Clearly, Model4 or Model5 is preferred, due to their much lower residual deviance (and misclassifcation rate, using 80% as the threshold of success). I would personally probably take Model4. It's simpler, and even though it has a higher misclassification rate, it has a lower rate of those predicted to live who actually die... this is the group I am most concerned about, so I would like to reduce the number of these errors.