BE7023 Homework 7

Mike Lape

November 7, 2018

```
#setwd("C:/Users/lapt3u/Box/UC/Fall_2018/BE7023_Adv_Biostats/adv_biostats/hw_7")
library(VGAM)
# Death < Vegetative < Major Disability < Minor Disability < Good Recovery
# Load the data
treat     <- c(1,2,3,4)
treat_names     <- c("Placebo", "Low", "Medium", "High")
death     <- c(59,48,44,43)
veg       <- c(25,21,14,4)
maj_dis   <- c(46,44,54,49)
min_dis   <- c(48,47,64,58)
good      <- c(32,30,31,41)

dat <- data.frame(treat, death, veg, maj_dis, min_dis, good)
total     <- rowSums(dat[2:6])
```

1. Postulate a multinomial logistic regression model by treating the covariate as numerical.

```
class(dat$treat)
```

```
## [1] "numeric"
```

```
mod <- vglm(cbind(death, veg, maj_dis, min_dis, good) ~ treat, data = dat,
            family = multinomial)
summary(mod)
```

```
##
## Call:
## vglm(formula = cbind(death, veg, maj_dis, min_dis, good) ~ treat,
##     family = multinomial, data = dat)
##
##
## Pearson residuals:
##   log(mu[,1]/mu[,5]) log(mu[,2]/mu[,5]) log(mu[,3]/mu[,5])
## 1            0.08942            -0.7686            -0.1321
## 2           -0.07082             1.0036            -0.1644
## 3           -0.27877             0.8917             0.6202
## 4            0.26667            -1.2268            -0.3458
##   log(mu[,4]/mu[,5])
## 1            -0.1301
## 2            -0.3331
## 3             0.9164
## 4            -0.4879
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  0.845882   0.277379   3.050 0.002292 **
```

```
## (Intercept):2  0.500403   0.348606   1.435 0.151161
## (Intercept):3  0.479402   0.283773   1.689 0.091145 .
## (Intercept):4  0.474027   0.279702   1.695 0.090121 .
## treat:1       -0.191477   0.101055  -1.895 0.058120 .
## treat:2       -0.545912   0.145031  -3.764 0.000167 ***
## treat:3       -0.044454   0.100906  -0.441 0.659541
## treat:4        0.003081   0.098705   0.031 0.975097
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  4
##
## Names of linear predictors:
## log(mu[,1]/mu[,5]), log(mu[,2]/mu[,5]), log(mu[,3]/mu[,5]), log(mu[,4]/mu[,5])
##
## Residual deviance: 5.9785 on 8 degrees of freedom
##
## Log-likelihood: -42.7708 on 8 degrees of freedom
##
## Number of iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
## Reference group is level  5  of the response
```

2. Count the number of parameters in the model.
   This model contains 8 parameters!

3. Fit the model to the data. Write the prediction model.
   The model was fitted in problem 1, the prediction equation is below.

$$Pr(Death|Treatment) = \frac{e^{(0.846-0.191*Treatment)}}{D}$$

$$Pr(VegState|Treatment) = \frac{e^{(0.500-0.546*Treatment)}}{D}$$

$$Pr(MajDisability|Treatment) = \frac{e^{(0.479-0.044*Treatment)}}{D}$$

$$Pr(MinDisability|Treatment) = \frac{e^{(0.474+0.003*Treatment)}}{D}$$

$$Pr(Good|Treatment) = \frac{1}{D}$$

$$D = 1 + e^{(0.846 - 0.191 * Treatment)} + e^{(0.500 - 0.546 * Treatment)} + e^{(0.479 - 0.044 * Treatment)} + e^{(0.474 + 0.003 * Treatment)}$$

4. Check the adequacy of the model

```
# Check the goodness of fit!
res_dev <- 5.9785
dof     <- 8
p       <- round(pchisq(res_dev, dof, lower.tail = F),3)

paste("p-val: ", p, sep = "")
```

```
## [1] "p-val: 0.65"
```

```
# Our p-value is 0.650, which means we cannot reject our null hypothesis and
# thus our multinomial model fits our data well.
```

5. Compare the observed and predicted probabilities and frequencies.

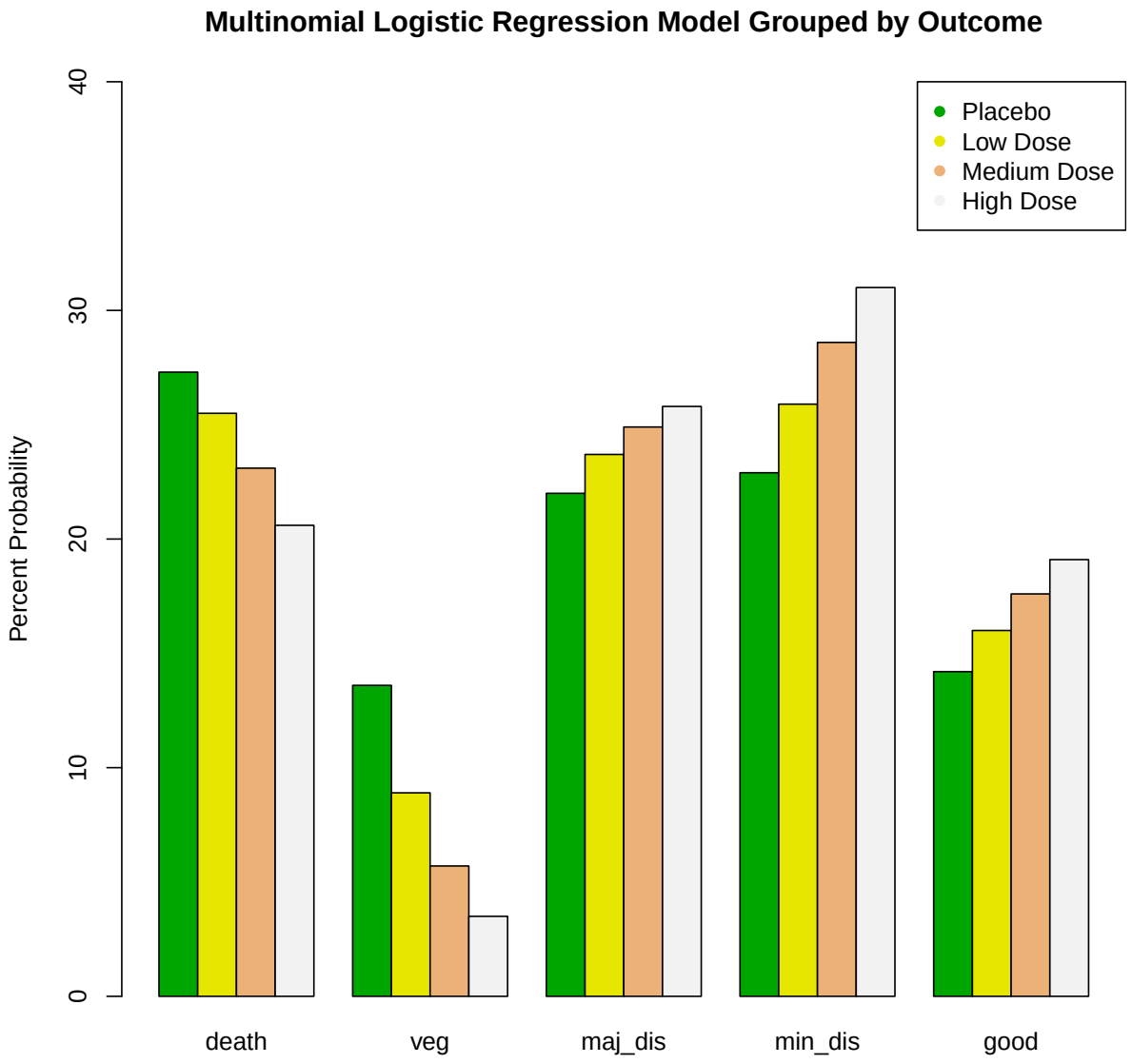
```
pred <- round(predict(mod, newdata = dat, type = "response"),3)
exp_freq  <- round(pred * total, 1)
comp <- data.frame(dat, exp_freq)

#comp

# I think it is easier to view the difference between our observed frequencies
# and expected frequencies directly, in which we can see that the
# predicted values agree quite well with the observed data
dat[,2:6] - exp_freq
```

```
##   death  veg maj_dis min_dis good
## 1   1.7 -3.6    -0.2    -0.1  2.2
## 2  -0.5  4.1    -1.0    -2.2 -0.4
## 3  -3.8  2.2     2.5     4.8 -5.4
## 4   2.8 -2.8    -1.3    -2.5  3.8
```

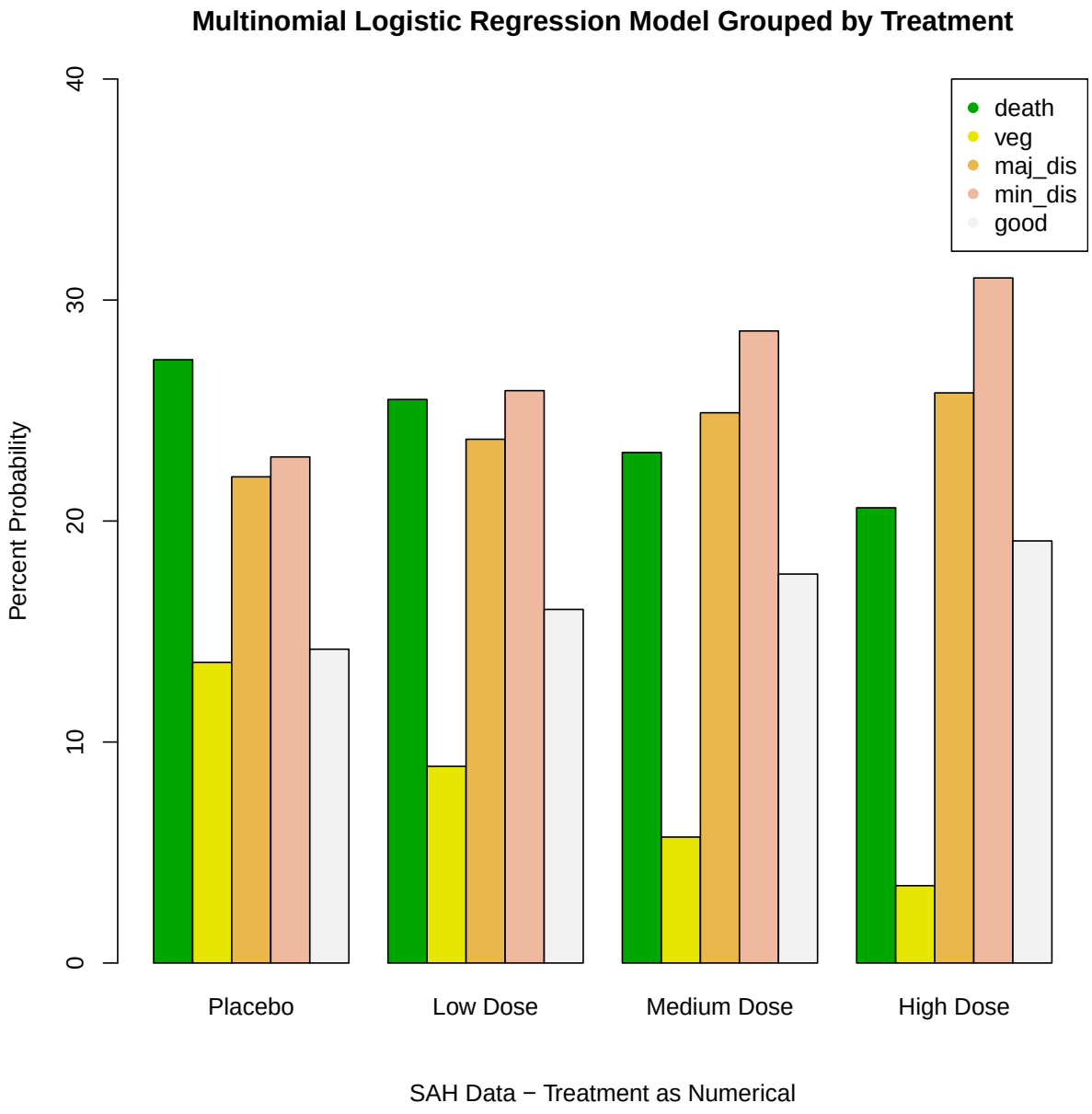
```
pred_perc <- pred * 100
rownames(pred_perc)  <- c("Placebo", "Low Dose", "Medium Dose", "High Dose")
```

6. Obtain bar plots of predicted probabilities in two different ways.

```
barplot(pred_perc, beside = T, col = terrain.colors(4), ylim = range(0,40))
title(ylab = "Percent Probability", main = "Multinomial Logistic Regression Model Grouped by Outcome",
      sub = "SAH Data - Treatment as Numerical")
legend(legend = rownames(pred_perc), pch = rep(16,4), col = terrain.colors(4), "topright")
```




```
# Transpose
barplot(t(pred_perc), beside = T, col = terrain.colors(5), ylim = range(0,40))
title(ylab = "Percent Probability", main = "Multinomial Logistic Regression Model Grouped by Treatment"
      sub = "SAH Data - Treatment as Numerical")
legend(legend = colnames(pred_perc), pch = rep(16,4), col = terrain.colors(5), "topright")
```



7. Postulate a multinomial logistic regression model by treating the covariate as categorical.

```
cat_dat       <- dat
cat_dat$treat <- treat_names
cat_dat$treat <- factor(cat_dat$treat)
class(cat_dat$treat)
```

```
## [1] "factor"
```

```
cat_mod <- vglm(cbind(death, veg, maj_dis, min_dis, good) ~ treat, data = cat_dat,
            family = multinomial, half.stepsizing = F)
summary(cat_mod)
```

```
##
```

```
## Call:
## vglm(formula = cbind(death, veg, maj_dis, min_dis, good) ~ treat,
##     family = multinomial, data = cat_dat, half.stepsizing = F)
##
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept):1   0.04763    0.21828   0.218 0.827275
## (Intercept):2  -2.32728    0.52382  -4.443 8.88e-06 ***
## (Intercept):3   0.17825    0.21166   0.842 0.399699
## (Intercept):4   0.34687    0.20404   1.700 0.089126 .
## treatLow:1      0.42238    0.31908   1.324 0.185594
## treatLow:2      1.97060    0.59611   3.306 0.000947 ***
## treatLow:3      0.20474    0.31758   0.645 0.519125
## treatLow:4      0.10208    0.31023   0.329 0.742121
## treatMedium:1   0.30257    0.32036   0.944 0.344925
## treatMedium:2   1.53235    0.61488   2.492 0.012699 *
## treatMedium:3   0.37675    0.30915   1.219 0.222976
## treatMedium:4   0.37802    0.29919   1.263 0.206412
## treatPlacebo:1  0.56417    0.30959   1.822 0.068405 .
## treatPlacebo:2  2.08042    0.58791   3.539 0.000402 ***
## treatPlacebo:3  0.18466    0.31271   0.591 0.554851
## treatPlacebo:4  0.05859    0.30613   0.191 0.848209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  4
##
## Names of linear predictors:
## log(mu[,1]/mu[,5]), log(mu[,2]/mu[,5]), log(mu[,3]/mu[,5]), log(mu[,4]/mu[,5])
##
## Residual deviance: -4.363e-14 on 0 degrees of freedom
##
## Log-likelihood: -39.7816 on 0 degrees of freedom
##
## Number of iterations: 7
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):2'
##
## Reference group is level  5  of the response
```

8. Count the number of parameters in the model.
   This model has 16 parameters!

9. Fit the model to the data. Write the prediction model.
   The model was fitted in problem 7, the prediction model can be found below.

$$Pr(Death|Treatment) = \frac{e^{(0.048+(0.422*Low_Dose)+(0.303*Mid_Dose)+(0.564*Placebo)}}{D}$$

$$Pr(VegState|Treatment) = \frac{e^{(-2.327+(1.971*Low_Dose)+(1.532*Mid_Dose)+(2.080*Placebo)}}{D}$$

$$Pr(MajDisability|Treatment) = \frac{e^{(0.178+(0.205*Low_Dose)+(0.377*Mid_Dose)+(0.185*Placebo)}}{D}$$

$$Pr(MinDisability|Treatment) = \frac{e^{(0.347+(0.102*Low_Dose)+(0.378*Mid_Dose)+(0.059*Placebo)}}{D}$$

$$Pr(Good|Treatment) = \frac{1}{D}$$

D = 1 + e^{0.048 + (0.422 * Low_Dose) + (0.303 * Mid_Dose) + (0.564 * Placebo)} + e^{-2.327 + (1.971 * Low_Dose) + (1.532 * Mid_Dose) + (2.080 * Placebo)} + e^{(0.178 + (0.205 * Low_Dose) + (0.377 * Mid_Dose) + (0.185 * Placebo)} + e^{(0.347 + (0.102 * Low_Dose) + (0.378 * Mid_Dose) + (0.059 * Placebo)}

High Dose (4) held as baseline for treatments.

10. Check the adequacy of the model.

```
# Check the goodness of fit!
cat_res_dev <- -4.363e-14
cat_dof     <- 0
cat_p       <- round(pchisq(cat_res_dev, cat_dof, lower.tail = F),3)

paste("p-val: ", cat_p, sep = "")
```

```
## [1] "p-val: 1"
```

```
# Our p-value is 1, which means we cannot reject our null hypothesis and
# thus our multinomial model fits our data well.
```

11. Compare the observed and predicted probabilities and frequencies.

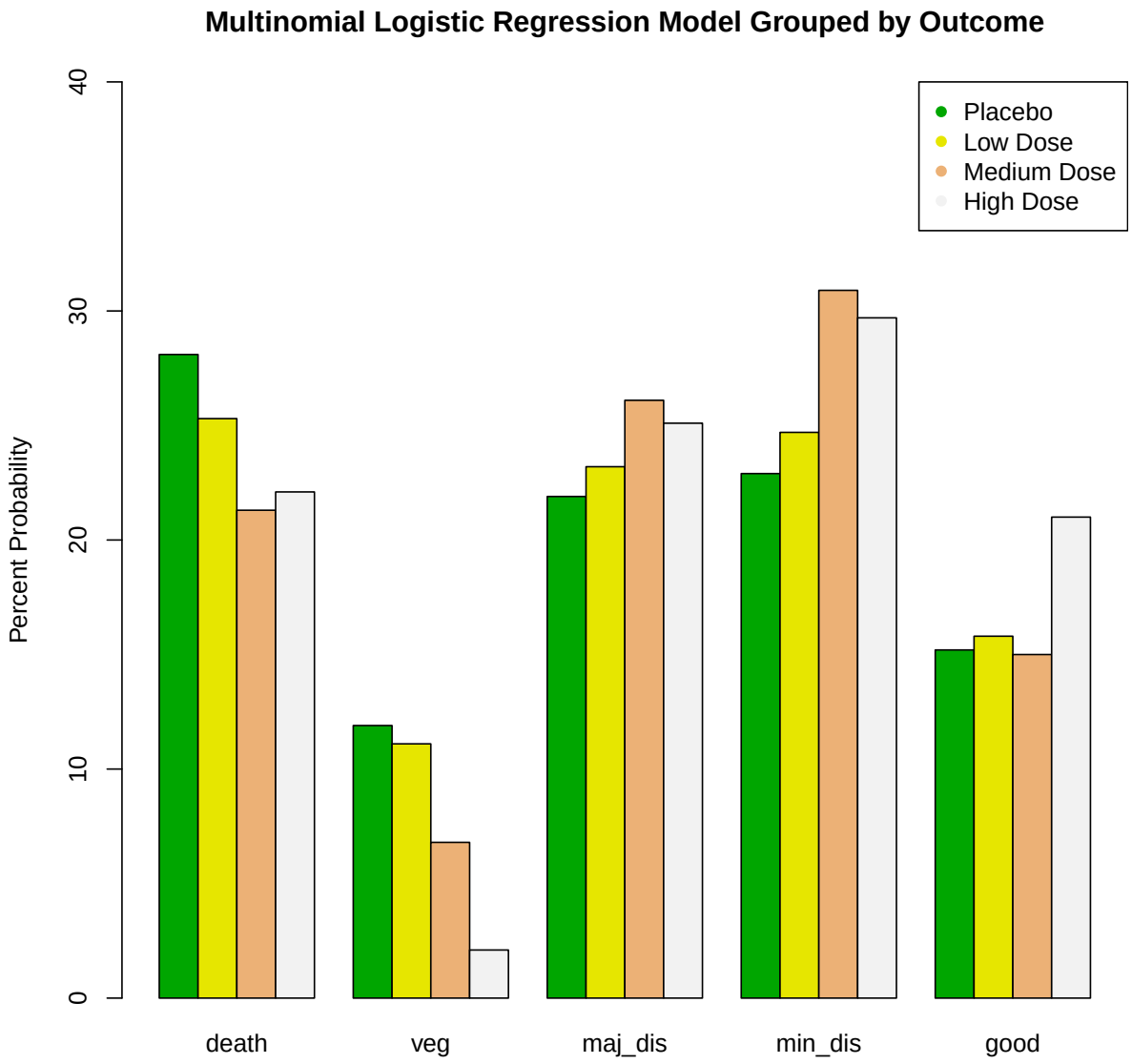
```
cat_pred <- round(predict(cat_mod, newdata = cat_dat, type = "response"),3)
cat_exp_freq  <- round(cat_pred * total, 1)
cat_comp <- data.frame(cat_dat, cat_exp_freq)

#cat_comp

# Comparing our observed frequencies and expected frequencies directly,
# in which we can see that the predicted values are almost exactly the
# observed data, much better than seen in the model where treatment was
# considered as a numeric.
cat_dat[,2:6] - cat_exp_freq
```

```
##   death  veg maj_dis min_dis good
## 1   0.0  0.0     0.0    -0.1  0.1
## 2  -0.1 -0.1    -0.1     0.1  0.0
## 3  -0.1 -0.1     0.0     0.0  0.0
## 4  -0.1 -0.1     0.1     0.1  0.1
```

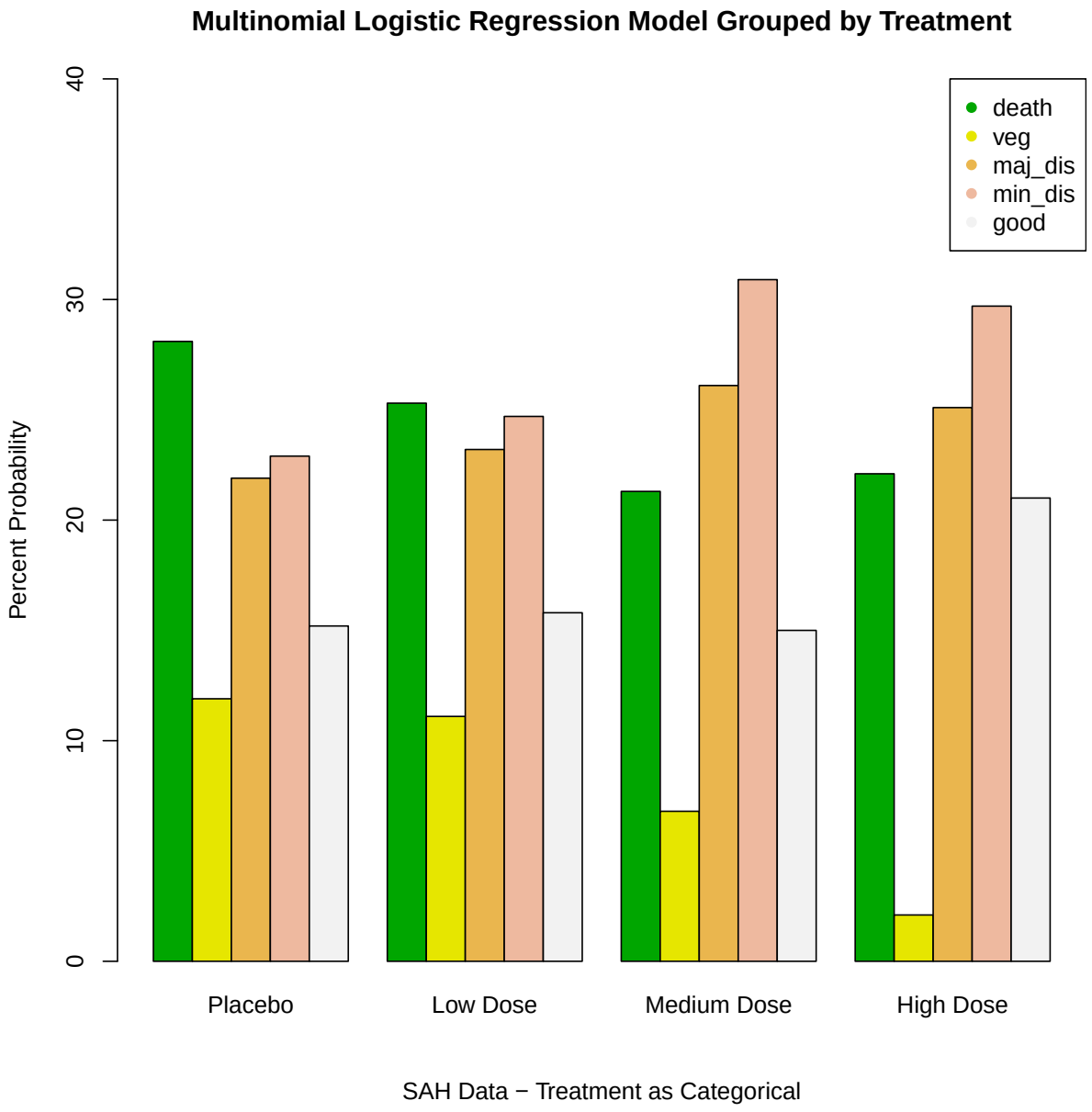
```
cat_pred_perc <- cat_pred * 100
rownames(cat_pred_perc)  <- c("Placebo", "Low Dose", "Medium Dose", "High Dose")
```

12. Obtain bar plots of predicted probabilities in two different ways

```
barplot(cat_pred_perc, beside = T, col = terrain.colors(4), ylim = range(0,40))
title(ylab = "Percent Probability", main = "Multinomial Logistic Regression Model Grouped by Outcome",
      sub = "SAH Data - Treatment as Categorical")
legend(legend = rownames(cat_pred_perc), pch = rep(16,4), col = terrain.colors(4), "topright")
```




```
# Transpose
barplot(t(cat_pred_perc), beside = T, col = terrain.colors(5), ylim = range(0,40),
        ylab = "Percent Probability", main = "Multinomial Logistic Regression Model Grouped by Treatmen
        sub = "SAH Data - Treatment as Categorical")
legend(legend = colnames(cat_pred_perc), pch = rep(16,4), col = terrain.colors(5), "topright")
```



13. Postulate a proportional odds model by treating the covariate as numerical.

```
class(dat$treat)
```

```
## [1] "numeric"
```

```
prop_mod <- vglm(cbind(death, veg, maj_dis, min_dis, good) ~ treat, data = dat,
            family = cumulative(par = T))

summary(prop_mod)
```

```
##
## Call:
## vglm(formula = cbind(death, veg, maj_dis, min_dis, good) ~ treat,
```

```
##     family = cumulative(par = T), data = dat)
##
##
## Pearson residuals:
##   logit(P[Y<=1]) logit(P[Y<=2]) logit(P[Y<=3]) logit(P[Y<=4])
## 1        -0.8742         1.4325       -0.17949        -0.9222
## 2        -0.6845         1.2073        0.19701        -0.2754
## 3        -0.1411        -0.8381       -0.08163         1.2159
## 4         1.9782        -1.9899        0.04382        -0.1565
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -0.71917    0.15881  -4.528 5.94e-06 ***
## (Intercept):2 -0.31860    0.15642  -2.037  0.04167 *
## (Intercept):3  0.69165    0.15793   4.380 1.19e-05 ***
## (Intercept):4  2.05700    0.17369  11.843  < 2e-16 ***
## treat         -0.17548    0.05632  -3.116  0.00183 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  4
##
## Names of linear predictors:
## logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3]), logit(P[Y<=4])
##
## Residual deviance: 18.1825 on 11 degrees of freedom
##
## Log-likelihood: -48.8728 on 11 degrees of freedom
##
## Number of iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
##     treat
## 0.8390501
```

14. Count the number of parameters in the model.
    This model has only 5 parameters in it.

15. Fit the model to the data. Write the prediction model.
    The model was fit in number 13, the prediction models can be found below.

$$Pr(Death|Treatment) = \frac{e^{(-0.719-0.175*treatment)}}{1+e^{(-0.719-0.175*treatment)}}$$

$$Pr(Death) + Pr(VegState) = \frac{e^{(-0.319-0.175*treatment)}}{1+e^{(-0.319-0.175*treatment)}}$$

$$Pr(Death) + Pr(VegState) + Pr(MajDisability) = \frac{e^{(0.692-0.175*treatment)}}{1+e^{(0.692-0.175*treatment)}}$$

$$Pr(Death) + Pr(VegState) + Pr(MajDisability) + Pr(MinDisability) = \frac{e^{(2.057-0.175*treatment)}}{1+e^{(2.057-0.175*treatment)}}$$

$$Pr(Death)+Pr(VegState)+Pr(MajDisability)+Pr(MinDisability)+Pr(Good) = 1-Pr(MinDisability)$$

where treatment: Placebo = 1 Low Dose = 2 Medium Dose = 3 High Dose = 4

16. Check the adequacy of the model.

```
prop_res_dev   <- 18.1825
prop_dof       <- 11
prop_p         <- round(pchisq(prop_res_dev, prop_dof, lower.tail = F),3)
prop_p
```

```
## [1] 0.077
```

```
# Our p-value is 0.077 so we still cannot reject our null hypothesis and thus
# our model adequately fits our data but definitely not as well as the
# multinomial logistic regression seen above.
```

17. Compare the observed and predicted probabilities and frequencies.

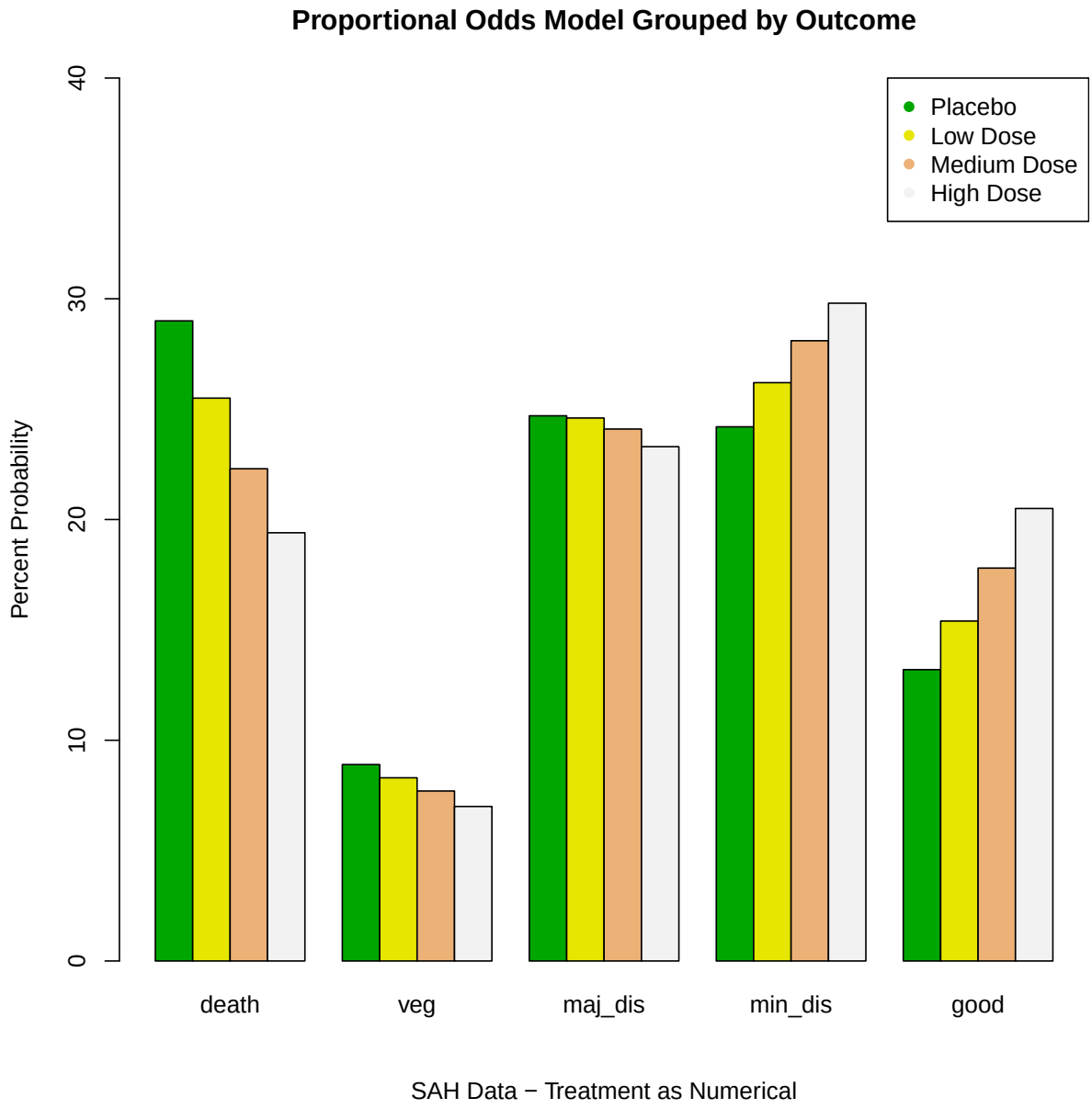
```
prop_pred <- round(predict(prop_mod, newdata = dat, type = "response"),3)
prop_exp_freq  <- round(prop_pred * total, 1)
prop_comp <- data.frame(dat, prop_exp_freq)

#prop_comp

# Comparing our observed frequencies and expected frequencies directly,
# in which we can see that there is more distance between the predicted
# values and the observed data, worse than both multinomial models.

# considered as a numeric.
dat[,2:6] - prop_exp_freq
```

```
##   death  veg maj_dis min_dis good
## 1  -1.9  6.3    -5.9    -2.8  4.3
## 2  -0.5  5.2    -2.7    -2.8  0.7
## 3  -2.2 -1.9     4.1     5.8 -5.8
## 4   5.2 -9.7     3.6    -0.1  1.0
```

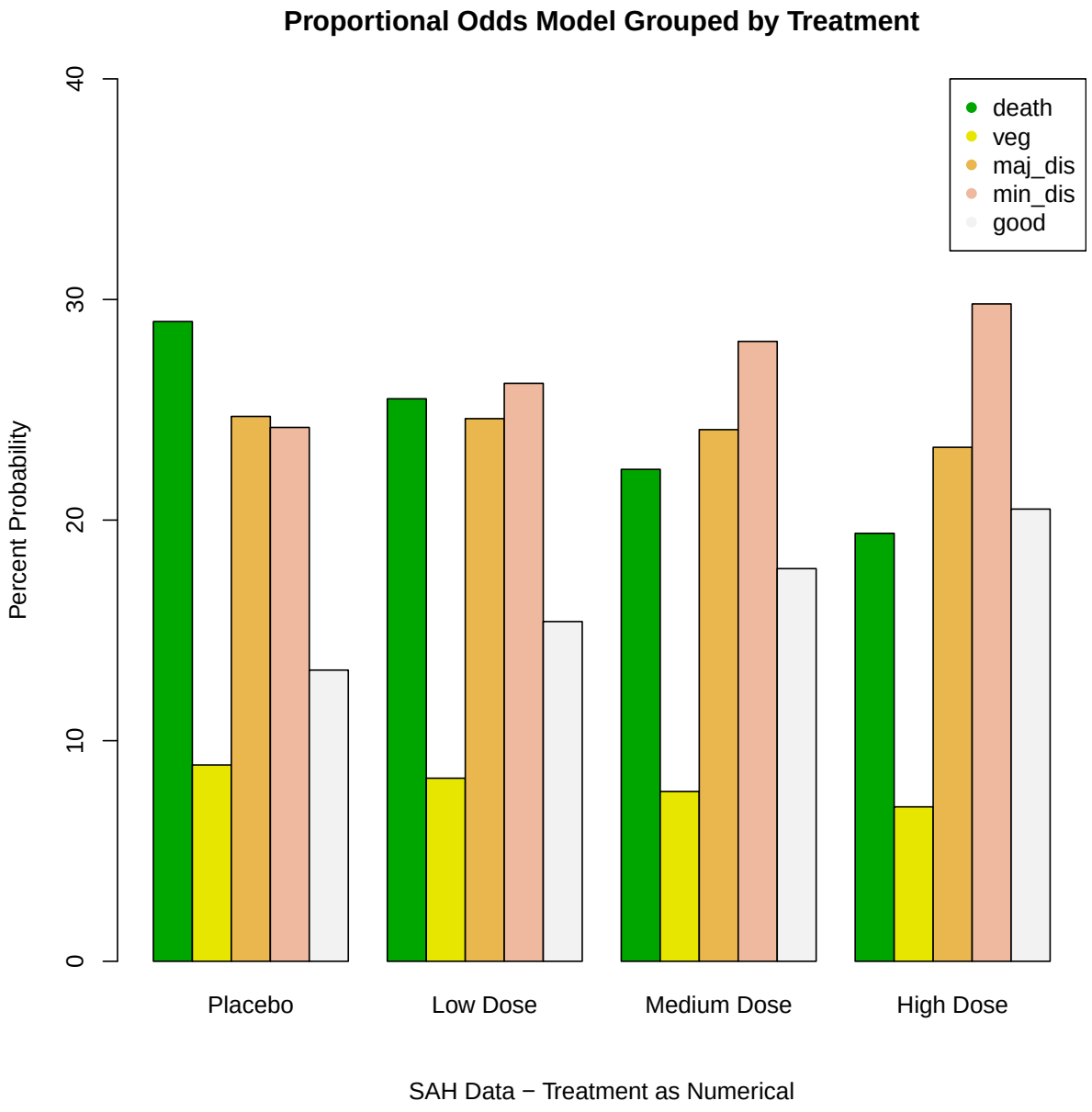
```
prop_pred_perc <- prop_pred * 100
rownames(prop_pred_perc)  <- c("Placebo", "Low Dose", "Medium Dose", "High Dose")
```

18. Obtain bar plots of predicted probabilities in two different ways.

```
barplot(prop_pred_perc, beside = T, col = terrain.colors(4), ylim = range(0,40))
title(ylab = "Percent Probability", main = "Proportional Odds Model Grouped by Outcome",
      sub = "SAH Data - Treatment as Numerical")
legend(legend = rownames(prop_pred_perc), pch = rep(16,4), col = terrain.colors(4), "topright")
```




```
# Transpose
barplot(t(prop_pred_perc), beside = T, col = terrain.colors(5), ylim = range(0,40))
title(ylab = "Percent Probability", main = "Proportional Odds Model Grouped by Treatment",
      sub = "SAH Data - Treatment as Numerical")
legend(legend = colnames(prop_pred_perc), pch = rep(16,4), col = terrain.colors(5), "topright")
```



19. Postulate a proportional odds model by treating the covariate as categorical.

```
class(cat_dat$treat)
```

```
## [1] "factor"
```

```
prop_cat_mod <- vglm(cbind(death, veg, maj_dis, min_dis, good) ~ treat, data = cat_dat,
            family = cumulative(par = T))

summary(prop_cat_mod)
```

```
##
## Call:
## vglm(formula = cbind(death, veg, maj_dis, min_dis, good) ~ treat,
```

```
##     family = cumulative(par = T), data = cat_dat)
##
##
## Pearson residuals:
##   logit(P[Y<=1]) logit(P[Y<=2]) logit(P[Y<=3]) logit(P[Y<=4])
## 1        -0.7787         1.5370       -0.07351       -0.83571
## 2        -0.8050         1.0713        0.04993       -0.39678
## 3        -0.1742        -0.8727       -0.12910        1.17971
## 4         2.0479        -1.9343        0.12977       -0.08275
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -1.43957    0.14157 -10.168  < 2e-16 ***
## (Intercept):2 -1.03903    0.13686  -7.592 3.15e-14 ***
## (Intercept):3 -0.02862    0.13173  -0.217  0.82800
## (Intercept):4  1.33708    0.14276   9.366  < 2e-16 ***
## treatLow       0.40322    0.18201   2.215  0.02674 *
## treatMedium    0.20338    0.17787   1.143  0.25287
## treatPlacebo   0.52078    0.17795   2.927  0.00343 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  4
##
## Names of linear predictors:
## logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3]), logit(P[Y<=4])
##
## Residual deviance: 18.0496 on 9 degrees of freedom
##
## Log-likelihood: -48.8064 on 9 degrees of freedom
##
## Number of iterations: 5
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
##     treatLow  treatMedium treatPlacebo
##     1.496629     1.225532     1.683333
```

20. Count the number of parameters in the model.
    The model has 7 parameters in it.

21. Fit the model to the data. Write the prediction model.
    The model was fitted in problem 19, the prediction model can be found below.

$$Pr(Death|Treatment) = \frac{e^{(-1.440+(0.403*LowDosase)+(0.203*MediumDosage)+(0.521*Placebo))}}{1+e^{(-1.440+(0.403*LowDosase)+(0.203*MediumDosage)+(0.521*Placebo))}}$$

$$Pr(Death) + Pr(VegState) = \frac{e^{(-1.039+(0.403*LowDosase)+(0.203*MediumDosage)+(0.521*Placebo))}}{1+e^{(-1.039+(0.403*LowDosase)+(0.203*MediumDosage)+(0.521*Placebo))}}$$

$$Pr(Death)+Pr(VegState)+Pr(MajDisability) = \frac{e^{(-0.029+(0.403*LowDosase)+(0.203*MediumDosage)+(0.521*Placebo))}}{1+e^{(-0.029+(0.403*LowDosase)+(0.203*MediumDosage)+(0.521*Placebo))}}$$

$$Pr(Death) + Pr(VegState) + Pr(MajDisability) + Pr(MinDisability) =$$

$$\frac{e^{(1.337+(0.403*LowDosase)+(0.203*MediumDosage)+(0.521*Placebo))}}{1+e^{(1.337+(0.403*LowDosase)+(0.203*MediumDosage)+(0.521*Placebo))}}$$

$$Pr(Death)+Pr(VegState)+Pr(MajDisability)+Pr(MinDisability)+Pr(Good) = 1-Pr(MinDisability)$$

22. Check the adequacy of the model.

```
prop_cat_res_dev   <- 18.0496
prop_cat_dof       <- 9
prop_cat_p         <- round(pchisq(prop_cat_res_dev, prop_cat_dof, lower.tail = F),3)
prop_cat_p
```

```
## [1] 0.035
```

```
# Our p-value is 0.035 which is less than 0.05 so we reject the null hypothesis
# and say this proportional odds model does not fit our data well.
```

23. Compare the observed and predicted probabilities and frequencies.

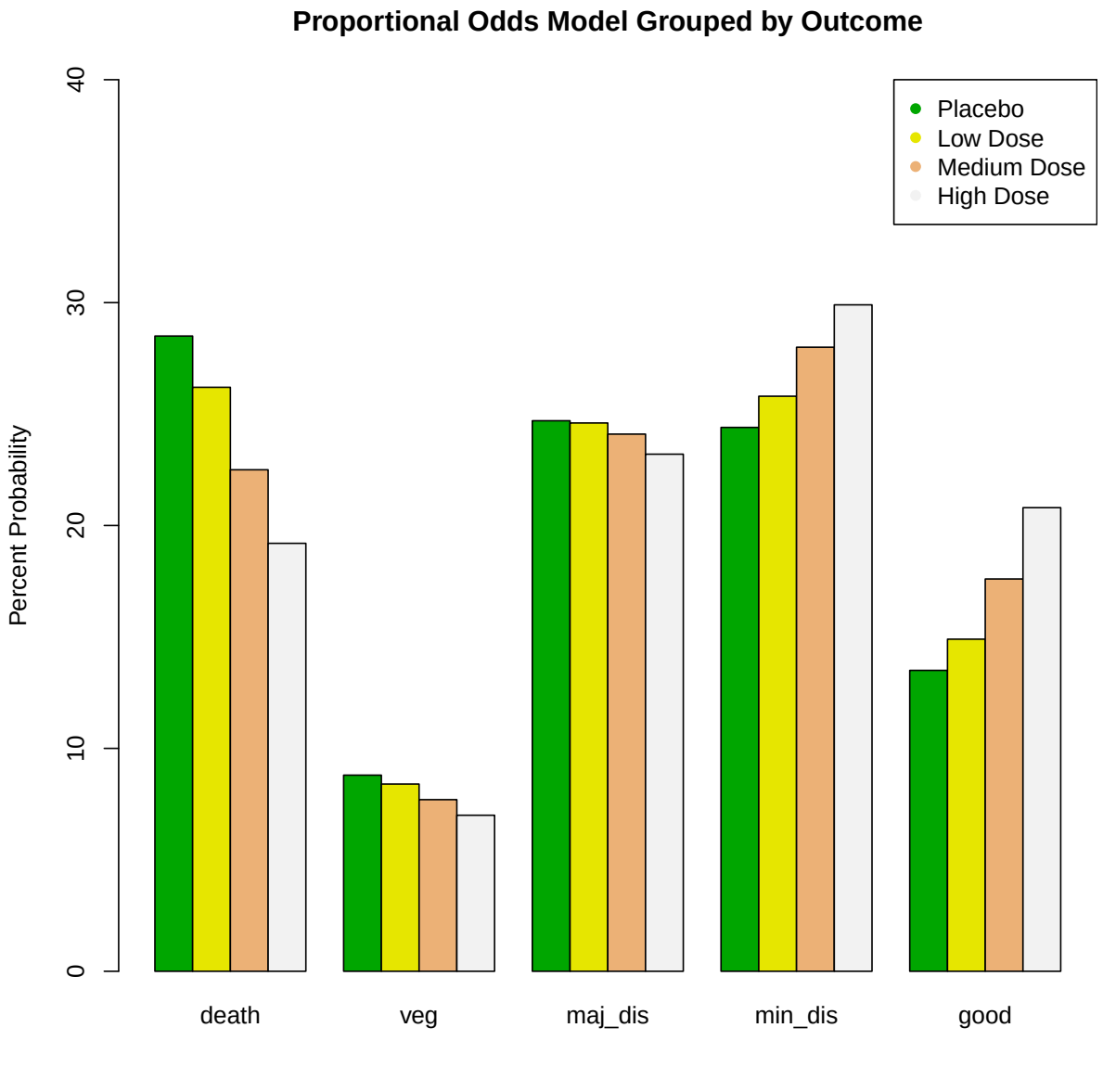
```
prop_cat_pred <- round(predict(prop_cat_mod, newdata = cat_dat, type = "response"),3)
prop_cat_exp_freq  <- round(prop_cat_pred * total, 1)
prop_cat_comp <- data.frame(cat_dat, prop_cat_exp_freq)

#prop_cat_comp

# Comparing our observed frequencies and expected frequencies directly,
# in which we can see that there is similar distance between the predicted
# values and the observed data as seen in the proportional odds model where
# treatment was treated as numerical and is still much worse than both multinomial
# models.
cat_dat[,2:6] - prop_cat_exp_freq
```

```
##   death  veg maj_dis min_dis good
## 1  -0.8  6.5    -5.9    -3.2  3.6
## 2  -1.8  5.0    -2.7    -2.0  1.7
## 3  -2.6 -1.9     4.1     6.0 -5.4
## 4   5.6 -9.7     3.8    -0.3  0.4
```

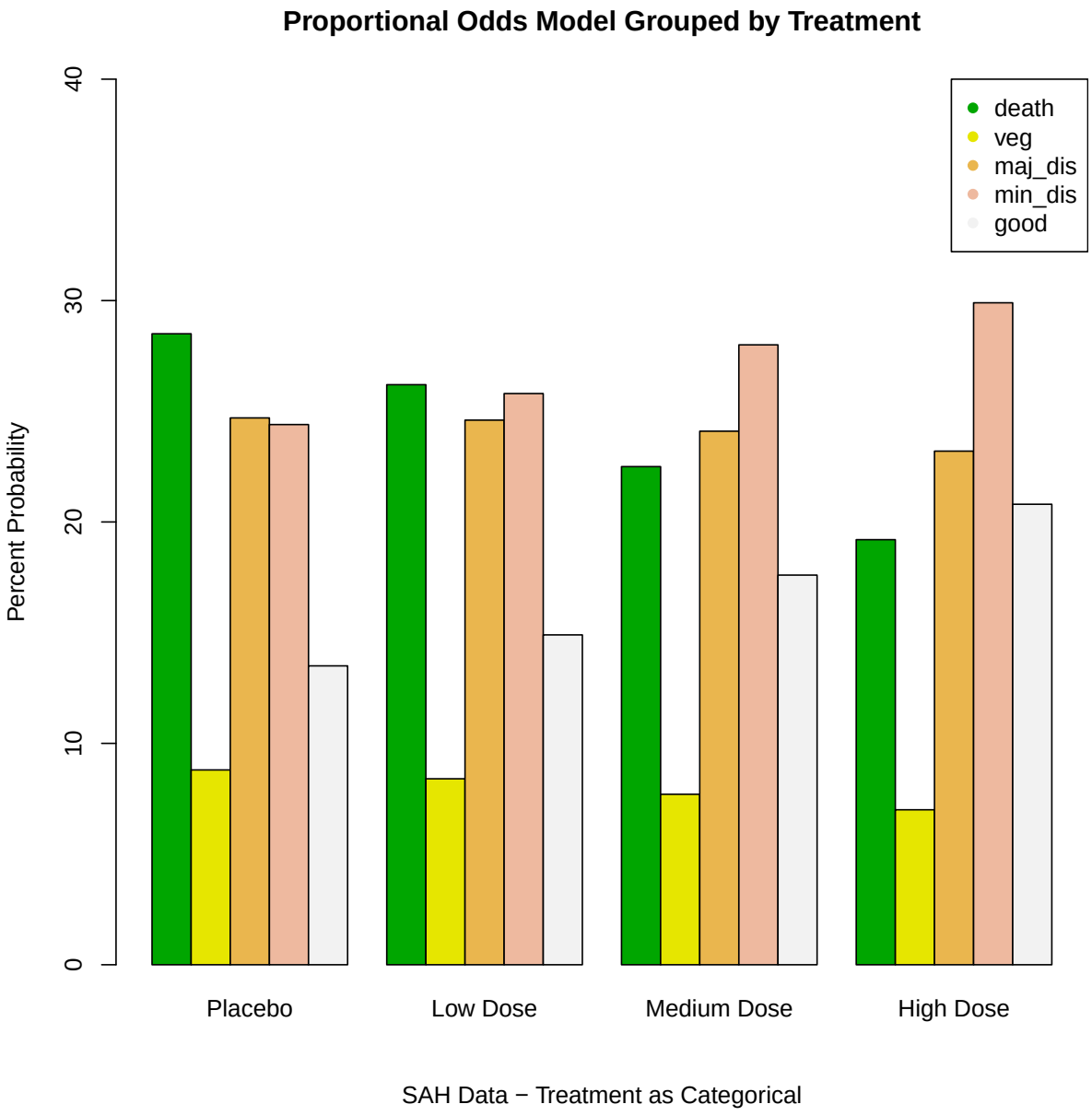
```r
prop_cat_pred_perc <- prop_cat_pred * 100
rownames(prop_cat_pred_perc)  <- c("Placebo", "Low Dose", "Medium Dose", "High Dose")
```

24. Obtain bar plots of predicted probabilities in two different ways.

```r
barplot(prop_cat_pred_perc, beside = T, col = terrain.colors(4), ylim = range(0,40))
title(ylab = "Percent Probability", main = "Proportional Odds Model Grouped by Outcome",
      sub = "SAH Data - Treatment as Categorical")
legend(legend = rownames(prop_cat_pred_perc), pch = rep(16,4), col = terrain.colors(4), "topright")
```




```
# Transpose
barplot(t(prop_cat_pred_perc), beside = T, col = terrain.colors(5), ylim = range(0,40))
title(ylab = "Percent Probability", main = "Proportional Odds Model Grouped by Treatment",
      sub = "SAH Data - Treatment as Categorical")
legend(legend = colnames(prop_cat_pred_perc), pch = rep(16,4), col = terrain.colors(5), "topright")
```



25. Compare and contrast the four models.
    Which of the four models would you propose as summary of the data?
    Why?
    Which treatment regimen would you recommend?
    Why?

Both of the proportional odds models have much worse p-values than the multinomial models, and the multinomial model where we consider treatment as a categorical has a p-value of 1. When comparing predicted frequencies and observed frequencies for the categorical multinomial model there is almost perfect agreement, which we don't see for any other models. And although the categorical multinomial model has more parameters at 16 the increased accuracy seems worth these extra parameters, so I would choose the categorical multinomial model. There is a decent chance this model is over-fit, but without any test data I cannot for sure tell that. I would recommend the high dosage as on the high dosage a patient has a slightly

higher chance of death than medium dose but has a much better probability of a good outcome. The high dose also gives the patient the smallest risk of a vegetative state outcome.